University of Pennsylvania
ScholarlyCommons

Publicly Accessible Penn Dissertations

1-1-2014

# Survival Analysis With Uncertain Endpoints Using an Internal Validation Subsample

Jarcy Zee
*University of Pennsylvania*, jarcy.zee@gmail.com

Follow this and additional works at: http://repository.upenn.edu/edissertations

 Part of the Biostatistics Commons

Recommended Citation
Zee, Jarcy, "Survival Analysis With Uncertain Endpoints Using an Internal Validation Subsample" (2014). *Publicly Accessible Penn Dissertations*. 1516.
http://repository.upenn.edu/edissertations/1516

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/edissertations/1516
For more information, please contact libraryrepository@pobox.upenn.edu.

# Survival Analysis With Uncertain Endpoints Using an Internal Validation Subsample

**Abstract**
When a true survival endpoint cannot be assessed for some subjects, an alternative endpoint that measures the true endpoint with error may be collected, which often occurs when the true endpoint is too invasive or costly to obtain. We develop nonparametric and semiparametric estimated likelihood functions that incorporate both uncertain endpoints available for all participants and true endpoints available for only a subset of participants. We propose maximum estimated likelihood estimators of the discrete survival function of time to the true endpoint and of a hazard ratio representing the effect of a binary or continuous covariate assuming a proportional hazards model. We show that the proposed estimators are consistent and asymptotically normal and develop the analytical forms of the variance estimators. Through extensive simulations, we also show that the proposed estimators have little bias compared to the naÃ¨ve estimator, which uses only uncertain endpoints, and are more efficient with moderate missingness compared to the complete-case estimator, which uses only available true endpoints. We illustrate the proposed method by estimating the risk of developing Alzheimer's disease using data from the Alzheimer's Disease Neuroimaging Initiative. Using our proposed semiparametric estimator, we develop optimal study design strategies to compare survival across treatment groups for a new trial with these data characteristics. We demonstrate how to calculate the optimal number of true events in the validation set with desired power using simulated data when assuming the baseline distribution of the true event, effect size, correlation between outcomes, and proportion of true outcomes that are missing can be estimated from pilot studies. We also propose a sample size formula that does not depend on baseline distribution of the true event and show that power calculated by the formula matches well with simulation based results. Using results from a Ginkgo Evaluation of Memory study, we calculate the number of true events in the validation set that would need to be observed for new studies comparing development of Alzheimer's disease among those with and without antihypertensive use, as well as the total number of subjects and number in the validation set to be recruited for these new trials.

**Degree Type**
Dissertation

**Degree Name**
Doctor of Philosophy (PhD)

**Graduate Group**
Epidemiology & Biostatistics

**First Advisor**
Sharon X. Xie

**Keywords**
Measurement Error, Missing Data, Study Design, Survival Analysis, Validation Sample

This dissertation is available at ScholarlyCommons: http://repository.upenn.edu/edissertations/1516

**Subject Categories**
Biostatistics

This dissertation is available at ScholarlyCommons: http://repository.upenn.edu/edissertations/1516

# SURVIVAL ANALYSIS WITH UNCERTAIN ENDPOINTS USING AN INTERNAL VALIDATION SUBSAMPLE

Jarcy Zee

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2014

Supervisor of Dissertation

_____

Sharon X. Xie

Associate Professor of Biostatistics

Graduate Group Chairperson

_____

John H. Holmes, Professor of Medical Informatics in Epidemiology

Dissertation Committee

Warren B. Bilker, Professor of Biostatistics

Susan S. Ellenberg, Professor of Biostatistics

Murray Grossman, Professor of Neurology

SURVIVAL ANALYSIS WITH UNCERTAIN ENDPOINTS USING AN INTERNAL VALIDATION SUBSAMPLE

© COPYRIGHT

2014

Jarcy Zee

This work is licensed under the

Creative Commons Attribution

NonCommercial-ShareAlike 3.0

License

To view a copy of this license, visit

http://creativecommons.org/licenses/by-nc-sa/3.0/

# ACKNOWLEDGEMENT

I would like to express my deepest appreciation to my dissertation advisor, Dr. Sharon X. Xie, for her dedication, patience, and guidance throughout the process of completing this thesis. I could not have asked for a more knowledgeable, more supportive, or kinder mentor and teacher who devoted an incredible amount of time and energy not only into my dissertation research, but also into helping me to become a successful biostatistician. I would also like to thank my committee members, Dr. Warren B. Bilker, Dr. Susan S. Ellenberg, and Dr. Murray Grossman for their insightful feedback that greatly improved my dissertation, as well as all of the collaborative research opportunities that they have provided for me.

Data collection and sharing for this dissertation was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and also supported by NIH grants P30 AG010129 and K01 AG030514. I would like to acknowledge my funding from NIH National Institute of Mental Health grant T32MH065218 and support for Dr. Sharon X. Xie from the NIH National Institute on Aging grant AG10124 (University of Pennsylvania Alzheimer's Disease Core Center).

I also thank all of the faculty, staff, and students of the Division of Biostatistics. In particular, I am grateful to my research advisor, Dr. Knashawn H. Morales, and master's thesis advisor, Dr. Mary D. Sammel, who taught me how to be an effective collaborative researcher and provided me with countless pieces of advice. I would also like to acknowledge Dr. Justine Shults, without whom I would not have become a PhD student at Penn and would not have found my first job after graduation.

Finally, I would like to recognize the role of my family and friends in helping me to complete this dissertation. I thank Parag Mahajan for his encouragement and for always being there for me every step of the way. I thank my sister, Dr. Tiffany Zee, for teaching me, sharing her experiences, and for her positivity. Most of all, I would like to thank my parents, Elena Chen and Samuel L. Zee, for their undying support and for all of the sacrifices that they have made for me throughout my entire life. Without their devotion, this dissertation would not be possible.

# ABSTRACT

SURVIVAL ANALYSIS WITH UNCERTAIN ENDPOINTS USING AN INTERNAL VALIDATION SUBSAMPLE

Jarcy Zee

Sharon X. Xie

When a true survival endpoint cannot be assessed for some subjects, an alternative endpoint that measures the true endpoint with error may be collected, which often occurs when the true endpoint is too invasive or costly to obtain. We develop nonparametric and semiparametric estimated likelihood functions that incorporate both uncertain endpoints available for all participants and true endpoints available for only a subset of participants. We propose maximum estimated likelihood estimators of the discrete survival function of time to the true endpoint and of a hazard ratio representing the effect of a binary or continuous covariate assuming a proportional hazards model. We show that the proposed estimators are consistent and asymptotically normal and develop the analytical forms of the variance estimators. Through extensive simulations, we also show that the proposed estimators have little bias compared to the naïve estimator, which uses only uncertain endpoints, and are more efficient with moderate missingness compared to the complete-case estimator, which uses only available true endpoints. We illustrate the proposed method by estimating the risk of developing Alzheimer's disease using data from the Alzheimer's Disease Neuroimaging Initiative. Using our proposed semiparametric estimator, we develop optimal study design strategies to compare survival across treatment groups for a new trial with these data characteristics. We demonstrate how to calculate the optimal number of true events in the validation set with desired power using simulated data when assuming the baseline distribution of the true event, effect size, correlation between outcomes, and proportion of true outcomes that are missing can be estimated from pilot studies. We also propose a sample size formula that does not depend on baseline distribution of the true event and show that power calculated by the formula matches well with simulation based results. Using results from a Ginkgo Evaluation of Memory study, we calculate the number of true events in the validation set that would need to be observed for new studies comparing development of Alzheimer's disease among those with and without antihypertensive use, as well as the total number of subjects and number in the validation set to be recruited for these new trials.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF ILLUSTRATIONS

# CHAPTER 1

## INTRODUCTION

In many clinical trials and epidemiological studies, the outcome of interest is time to an event, such as disease progression. The true outcome in these studies is often too invasive or costly to obtain, but alternative outcomes measure the true outcome with error. For example, the gold standard method for assessing time to renal function halving measures glomerular filtration rate (GFR), which is expensive and cumbersome to patients, but using equations based on serum creatinine to estimate GFR is inexact (Stevens et al., 2006). Another example is in the time to pathological diagnosis of Alzheimer's disease (AD), which can be accurately obtained with a cerebral spinal fluid (CSF) assay of amyloid beta ($A\beta$) protein concentrations (Shaw et al., 2009), but the lumbar puncture required for the procedure is often considered too painful for patients. An alternative method of AD diagnosis more widely used in practice is based on evaluation of clinical symptoms and cognitive tests, but the symptoms of AD are often mistaken for other types of dementia (Jack Jr et al., 2010).

Sometimes, it is possible to obtain both the uncertain or mismeasured outcome on all patients and the true outcome on just a subset of patients. For these situations, it is important to develop powerful analytical approaches to use the combined data, since standard methods for conducting survival analysis utilize only one endpoint. This dissertation will develop innovative statistical methods to conduct analyses on discrete time-to-event data with these characteristics. The proposed approach can estimate survival functions and hazard ratios for the effects of binary or continuous covariates, and it is shown to be superior to standard approaches that use only one of the outcomes. Optimal study design strategies for designing new studies to implement the proposed method are also developed.

## 1.1. Background

### *1.1.1. Methods with Uncertain Endpoints*

When true outcomes are difficult to obtain, uncertain outcomes are often used as an alternative due to their wide availability. Standard survival analysis methods, such as the Kaplan-Meier estimate of the survival function or Cox proportional hazards model for assessing covariate effects, may give

biased results when using these uncertain outcomes. Several novel statistical methods have been proposed to address this issue, but many rely on prior knowledge of the mismeasurement rates of the uncertain endpoint.

Snapinn (1998) used weights representing certainty of potential endpoints to modify the Cox proportional hazards model. Each of these weights are based on posterior probabilities that the potential endpoint is a true endpoint, which rely either on assumptions about the characteristics of the diagnostic tools or on an "endpoint committee" who must estimate them. As Snapinn indicates, however, it may be difficult to obtain the appropriate weights accurately, which diminishes the method's performance (Snapinn, 1998).

Richardson and Hughes (2000) developed Expectation-Maximization (EM) algorithms for estimating the distribution of time to an event using uncertain outcomes. Meier, Richardson, and Hughes (2003) extended this work to a semiparametric version assuming a proportional hazards model. Both methods produce less biased survival estimates than standard survival analysis methods and use supplemented EM algorithms to estimate variance-covariance matrices of parameter estimates. Similarly, Balasubramanian and Lagakos (2001) assumed a known time-dependent sensitivity function to estimate the distribution of the time to perinatal HIV transmission. However, all of these methods rely on known rates of sensitivity and specificity of the diagnostic tool used to determine the uncertain outcome. These rates may not be available and estimates may not be accurate. Magaret (2008) showed that even a 2 percent inaccuracy of specificity can cause a 14.5 percent bias in parameter estimates.

### *1.1.2. Use of a Validation Subsample*

Although uncertain outcomes are mismeasured and can lead to biased parameter estimates, true outcomes may also be available for a subset of patients. Using standard analysis methods on just the true outcomes limits the sample size and therefore power in making inference. However, statistical methods have been developed for the situation where both uncertain outcomes are available on all patients and true outcomes available in a subset, known as a validation subsample. Specifically, Pepe (1992) developed an estimated likelihood method for these types of data, although not specifically for a survival setting.

For true outcome $Y$, uncertain outcome $S$, covariates, $Z$, and parameters of interest, $\beta$, Pepe's

estimated likelihood takes the following form:

$$\hat{L}(\beta) = \prod_{i \in V} P_\beta(Y_i|Z_i) \prod_{j \in \bar{V}} \int_y P_\beta(y|Z)\hat{P}(S|y,Z)dy \tag{1.1}$$

where $V$ represents the validation subsample, the set of patients who have both true and uncertain outcomes, $\bar{V}$ represents the non-validation set, in which patients only have uncertain outcomes, and $\hat{P}(S|y,Z)$ is estimated empirically (Pepe, 1992). Therefore, those who have the true outcomes contribute the probability distribution of the true outcomes, as in a standard likelihood. Those who only have the uncertain outcomes contribute an estimated probability distribution of the uncertain outcomes, which incorporates the relationship between the true and uncertain outcomes. This relationship is estimated by observing the values of the true and uncertain outcomes within the validation subsample.

Although Pepe's original work was not designed for survival outcomes, Fleming et al. (1994) used the estimated likelihood method for the proportional hazards model by incorporating a validation set available on all subjects (i.e., no missing true endpoint measures). In this special case, having the uncertain outcomes is only useful in augmenting the likelihood for subjects with censored true failure times. Magaret (2008) also extended Pepe's work to the discrete proportional hazards model for situations where outcomes are only validated when the uncertain event status is positive. Therefore, Magaret's method is useful for data with only a subsample of true outcomes, but it does not allow for any false negatives. In addition, the method assumes there are no missed visits, so only type 1 right censoring (i.e., censoring time is not random) is allowed. Finally, the method only considers discrete covariates of interest.

### *1.1.3. Study Design*

Several methods exist to compute sample size for standard survival analysis studies (Freedman, 1982; Lakatos, 1986, 1988; Schoenfeld, 1981, 1983; Shih, 1995). Freedman (1982) developed a sample size formula for the logrank test and compared the power calculated by the formula to that using Monte Carlo simulations to show that the formula worked well. Schoenfeld (1981; 1983) developed a similar formula for use with either a logrank test or a Cox proportional hazards model, derived by exploiting asymptotic properties of the corresponding score test statistics. Both sample size formulas are widely used for clinical trials using standard survival analysis methods, but these

methods are for single outcomes and do not take into account any potential mismeasurement.

## 1.2. Novel Developments

In this dissertation, we fill in the gaps in the literature by developing flexible methods for the design and analysis of time-to-event data with uncertain outcomes using a validation subsample. The dissertation consists of three parts. In Chapter 2, we first propose a nonparametric discrete survival function estimator. There are three new contributions to the literature from this paper which we summarize below. First, we propose a maximum estimated likelihood estimator that incorporates both uncertain outcomes on all subjects and true outcomes on a validation subsample without assuming known mismeasurement rates of the uncertain outcomes. We assume study subjects are evaluated at predetermined time points by study design, so survival time is a discrete random variable for both true and uncertain endpoints. Second, we allow for missingness of the true outcome regardless of the value of the uncertain event indicator, so both false negatives and false positives are allowed. Third, our proposed estimator is able to handle both type 1 and random right censoring mechanisms. We develop the asymptotic distribution theory for the proposed estimator and provide an asymptotic variance estimator. We also demonstrate the performance of our proposed estimator through extensive simulations and illustrate the use of the method by estimating the survival function for the time to AD diagnosis using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

In Chapter 3, we develop a semiparametric approach to estimate a hazard ratio representing the effect of a covariate of interest assuming a proportional hazards model. In addition to the contributions to the literature above, the method discussed in this chapter allows for either a binary or a continuous covariate. Unlike in previous literature and in the nonparametric version, the estimated likelihood that incorporates a continuous covariate requires the use of a smooth kernel-type function in estimation. For both the binary and continuous covariate cases, we develop the asymptotic theory for the estimate of the log hazard ratio and its asymptotic variance estimator while treating all other parameters as nuisance parameters. We test the semiparametric estimated likelihood method using extensive simulations. Finally, using the ADNI data, we illustrate the method by estimating the effect of gender (binary) and years of education (continuous) on time to AD diagnosis.

In Chapter 4, we develop study design strategies to find the optimal number of true events in the

validation subsample when using the semiparametric estimated likelihood method to assess treatment effects. We calculate sample sizes assuming the goal is to achieve a pre-specified power for a Wald-type test to detect differences between treatment groups. We develop optimal designs based on simulations for a range of study conditions, including varying effect sizes, correlations between outcomes, percentage of missing true outcomes, number of time points in the study, and baseline distributions. We also propose the use of a sample size formula adapted from Schoenfeld's (1983) formula for the Cox proportional hazards model and demonstrate its performance by comparing calculated power to those from Monte Carlo simulations. Finally, we conclude in Chapter 5 and discuss future directions of study.

# CHAPTER 2

# NONPARAMETRIC DISCRETE SURVIVAL FUNCTION ESTIMATION WITH UNCERTAIN ENDPOINTS USING AN INTERNAL VALIDATION SUBSAMPLE

## 2.1. Introduction

Survival function estimation is crucial in studying disease progression and therapeutic benefits of drugs in epidemiology studies and clinical trials that involve time-to-event data. However, event outcomes may be subject to measurement error, which can lead to misclassification of the true event outcome. Gold standard or better outcome measurements are sometimes unavailable due to high costs or invasive procedures, and using only complete, true outcomes may exclude many subjects due to missing data. For example, the pathological diagnosis of Alzheimer's disease (AD) has been traditionally determined by autopsy. Recently, as we enter the exciting new era of "personalized medicine," AD biomarker research has been very successful. It is well accepted now that time to pathological diagnosis of AD can be reliably measured by time to an abnormal biomarker value among living participants in research studies (Shaw et al., 2009). Specifically, the amyloid beta ($A\beta$) protein biomarker from a cerebral spinal fluid (CSF) assay has been shown to represent the pathological aspects of AD well and the abnormality of $A\beta$ can be used as a reliable (true) endpoint for studying time to pathological diagnosis of AD among living participants (Shaw et al., 2009). However, the CSF biomarker assay involves a lumbar puncture, so it is often considered too invasive for many patients and therefore has limited availability. An alternative outcome is time to diagnosis of AD by clinical assessment, which relies primarily on cognitive tests. The clinical diagnosis is widely available, but it measures the outcome of pathological diagnosis with error. Sources of error in clinical diagnosis include normal aging independent of AD, "cognitive reserve" due to education-linked factors, and disease heterogeneity (Nelson et al., 2012). Thus, the clinical diagnosis is an uncertain endpoint. Under these circumstances, it is important to develop powerful analytical approaches to use combined information from both true and uncertain endpoints to obtain consistent and more efficient estimators compared to the naïve estimator, which ignores true endpoint measures, and the complete-case estimator, which uses only the available true endpoint measures.

Our proposed method is motivated by survival function estimation of time to pathological development of AD using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Weiner et al., 2012). Participants in the ongoing ADNI study were evaluated at predetermined time points to assess AD development based on cognitive tests. Regardless of these clinical diagnoses, a subset of participants also had longitudinal CSF assays to measure $A\beta$ values, from which time to CSF diagnoses could be determined. Some study participants randomly withdrew from the study before developing cognitive or pathological signs of AD. Therefore, survival time is a discrete random variable subject to random right censoring. Although several nonparametric and semiparametric methods for estimating survival outcomes when the outcome is uncertain have been proposed, many rely on prior knowledge of the mismeasurement rates of the uncertain endpoint without an internal validation subsample of true endpoints (Snapinn, 1998; Richardson and Hughes, 2000; Meier et al., 2003; Balasubramanian and Lagakos, 2001). Among those that do incorporate a validation subsample, the method primarily focused on the discrete proportional hazards model requiring that validation is only performed on those with positive uncertain endpoints, and the method cannot handle the random censoring we have in the ADNI data (Magaret, 2008).

Specifically, Snapinn (1998) estimated weights representing certainty of potential endpoints to modify the Cox proportional hazards model. Richardson and Hughes (2000) obtained unbiased product limit estimates of time to an event when the event indicator has measurement error using an Expectation-Maximization (EM) algorithm. Their estimate uses known information about the sensitivity and specificity of the diagnostic test for having the event without a validation sample. Meier, Richardson, and Hughes (2003) extended this work for the adjusted proportional hazards model for discrete failure times, also assuming known sensitivity and specificity. Similarly, Balasubramanian and Lagakos (2001) assumed a known time-dependent sensitivity function to estimate the distribution of the time to perinatal HIV transmission.

Pepe (1992) developed an estimated likelihood method to incorporate both uncertain endpoints and a validation subsample to make inference without assuming known sensitivity or specificity, but not specifically for a survival setting. Fleming et al. (1994) used Pepe's method for the proportional hazards model by incorporating a validation set available on all subjects (i.e., no missing true endpoint measures) to augment the likelihood for subjects with censored failure times. Magaret (2008) also extended Pepe's work to the discrete proportional hazards model in a method designed for

situations where outcomes were only validated when the mismeasured event status was positive, so false-negatives were not possible. Because the method assumes no missed visits, only type 1 right censoring (i.e., censoring time is not random) is allowed. Therefore, these previous methods are unable to address the unique challenges seen in the ADNI data.

We propose a nonparametric discrete survival function estimator for data with characteristics similar to those of the ADNI study. There are three new contributions to the literature from this paper which we summarize below. First, we propose a nonparametric discrete survival function estimator without assuming known mismeasurement rates of the uncertain outcome. Instead, we incorporate information from an internal validation subsample to construct the survival function estimator. We use Pepe's (1992) framework to construct an estimated likelihood for the survival function of time to an event, incorporating both an uncertain observed time and event indicator on all subjects and a true observed time and event indicator on a validation subsample. The proposed estimator is the nonparametric maximum estimated likelihood survival function estimator. In addition, because study subjects are evaluated at predetermined time points by study design, survival time is a discrete random variable for both true and uncertain endpoints. We develop the asymptotic distribution theory and provide an asymptotic variance estimator. Second, the proposed nonparametric survival function estimator allows missingness of the true endpoint regardless of the value of the uncertain event indicator. In other words, validation can be conducted on subjects with either observed or censored uncertain events. Third, the proposed estimator is able to handle both type 1 and random right censoring mechanisms. Our allowance of random censoring and objective of estimating an entire survival function provide some unique challenges in using survival outcomes as compared to Pepe's (1992) original work.

We organize the rest of the article as follows. We first describe the estimated likelihood and nonparametric maximum estimated likelihood estimator (Section 2.2). We then develop the asymptotic properties of the proposed estimator (Section 2.3). We perform extensive simulations to assess the performance of our proposed estimator and compare it to the complete-case and naïve Kaplan-Meier survival function estimators (Section 2.4). The simulations consider different correlations between true and uncertain endpoints, different amounts and types of censoring, as well as different amounts of missingness of true endpoints. This is followed by an application to the estimation of the survival function of time to pathological diagnosis of Alzheimer's disease using data from the

ongoing ADNI study (Section 2.5). Finally, we summarize our findings and point to applications where incorporating both true and uncertain endpoints are particularly useful (Section 2.6).

## 2.2. Proposed Nonparametric Maximum Estimated Likelihood Estimator

Let $T$ represent the true time to event and $C$ represent the true right censoring time, with event indicator $\delta = I(T \leq C)$. Similarly, let $T^*$ represent the uncertain time to event and $C^*$ be the uncertain right censoring time, with indicator $\delta^* = I(T^* \leq C^*)$. Define $X = \min\{T, C\}$ and $X^* = \min\{T^*, C^*\}$. Then $X$ and $X^*$ represent the true and uncertain observed times, respectively. Let $x_k$ represent the $k$th unique, ordered observed true time point for $k = 1, \cdots, K$, where $K$ is the total number of unique true observed times. Let $F$ represent the survival function of the true time to event and let $G$ represent the survival function of the true censoring time.

Let $V$ represent the validation set, where both the uncertain and true outcomes are available. There are $n_V$ subjects in the validation set. It is assumed that the validation subsample is a representative sample of the entire cohort, implying that data are missing completely at random. Then $\bar{V}$ is the non-validation set, where only the uncertain outcome is available and the true outcome is missing. With a total of $n$ subjects in the study, there are $n - n_V$ subjects in the non-validation set. The entire observed data are $(X_i, \delta_i, X_i^*, \delta_i^*)$ for $i = 1, \cdots, n_V$ and $(X_j^*, \delta_j^*)$ for $j = 1, \cdots, n - n_V$. Using similar arguments as in Pepe (1992), the full likelihood would then be

$$L = \prod_{i \in V} P(X_i, \delta_i) P(X_i^*, \delta_i^* | X_i, \delta_i) \prod_{j \in \bar{V}} P(X_j^*, \delta_j^*). \tag{2.1}$$

To avoid having to specify or assume the form of the relationship between the true and uncertain endpoints, we propose to use the estimated likelihood

$$\hat{L} = \prod_{i \in V} P(X_i, \delta_i) \hat{P}(X_i^*, \delta_i^* | X_i, \delta_i) \prod_{j \in \bar{V}} \hat{P}(X_j^*, \delta_j^*), \tag{2.2}$$

where for discrete data,

$$\hat{P}(X_j^*, \delta_j^*) = \sum_{k=1}^{K} \sum_{\delta=0}^{1} P(x_k, \delta) \hat{P}(X_j^*, \delta_j^* | x_k, \delta). \tag{2.3}$$

The sum marginalizes the joint distribution to obtain the marginal distribution of the uncertain out-

come, so the outer sum is taken over all possible time points, $k = 1, \cdots, K$. The estimated conditional probability $\hat{P}(X_j^*, \delta_j^* | x_k, \delta)$ is given by

$$\hat{P}(X_j^*, \delta_j^* | x_k, \delta) = \frac{\hat{P}(X_j^*, \delta_j^*, x_k, \delta)}{\hat{P}(x_k, \delta)} \tag{2.4}$$

$$= \frac{\frac{1}{n_V} \sum_{i \in V} I(X_i^* = X_j^*, \delta_i^* = \delta_j^*, X_i = x_k, \delta_i = \delta)}{\frac{1}{n_V} \sum_{i \in V} I(X_i = x_k, \delta_i = \delta)}, \tag{2.5}$$

where $I(\cdot)$ is the indicator function. Conceptually, the conditional probability is estimated empirically by counting the proportion of subjects in the validation set whose uncertain outcomes match those of the given non-validation set subject. Because the conditional probability $\hat{P}(X_i^*, \delta_i^* | X_i, \delta_i)$ from the validation set contribution does not contain any parameters, it can be factored out of the likelihood and the estimated likelihood to be maximized becomes

$$\hat{L} \propto \prod_{i \in V} P(X_i, \delta_i) \prod_{j \in \bar{V}} \hat{P}(X_j^*, \delta_j^*). \tag{2.6}$$

Then for a subject $i \in V$, the contribution to the likelihood is the same as it would be in a standard discrete survival setting,

$$P(X_i, \delta_i) = \{F(x_{k_i - 1}) - F(x_{k_i})\}^{\delta_i} F(x_{k_i})^{1 - \delta_i} G(x_{k_i - 1})^{\delta_i} \{G(x_{k_i - 1}) - G(x_{k_i})\}^{1 - \delta_i} \tag{2.7}$$

$$\propto \{F(x_{k_i - 1}) - F(x_{k_i})\}^{\delta_i} F(x_{k_i})^{1 - \delta_i} \tag{2.8}$$

where $x_{k_i}$ is the observed time for subject $i$ corresponding to the $k$th unique observed time point. Only the true outcome contributes to the likelihood for those in the validation set, implying that uncertain outcomes do not provide any additional information when the true outcome is known. However, the uncertain outcomes for those in the validation set are still used to estimate the relationship between the uncertain and true outcomes, which are then used to weight likelihood contributions for those in the non-validation set. For a subject $j \in \bar{V}$, the contribution to the likelihood is

$$\hat{P}(X_j^*, \delta_j^*) = \sum_{k=1}^{K} \sum_{\delta=0}^{1} \left[ \{F(x_{k-1}) - F(x_k)\}^{\delta} F(x_k)^{1-\delta} G(x_{k-1})^{\delta} \{G(x_{k-1}) - G(x_k)\}^{1-\delta} \cdot \right.$$

$$\left. \frac{\frac{1}{n_V} \sum_{i \in V} I(X_i^* = X_j^*, \delta_i^* = \delta_j^*, X_i = x_k, \delta_i = \delta)}{\frac{1}{n_V} \sum_{i \in V} I(X_i = x_k, \delta_i = \delta)} \right]. \tag{2.9}$$

Unlike in the validation set contribution, the censoring distribution cannot be factored out of the likelihood from the non-validation set contribution. This distribution is important in allowing random censoring for survival outcomes in the estimated likelihood method. Note that any subjects in the non-validation set with an observed uncertain time that does not match any observed uncertain times in the validation set do not contribute to the likelihood.

There are two special cases worth considering. First, in the situation where the uncertain outcome is perfect, or $P(X, \delta|X^*, \delta^*) = 1$, the likelihood reduces to that of the standard likelihood where all subjects have the true outcome. An example of this situation is when the uncertain outcome has no measurement error or is exactly the same as the true outcome. Second, in the situation where the uncertain outcome is useless, or $P(X^*, \delta^*|X, \delta) = P(X^*, \delta^*)$, the likelihood reduces to that of the standard likelihood where there is no non-validation set. Additional details on the derivation of the estimated likelihood and on these special cases are available in Appendix A.1.

The estimated likelihood is a function of possible survival function values for the event distribution and censoring distribution at each time point. The parameters representing the censoring distribution $G$ are estimated jointly with the parameters representing the event distribution $F$, but treated as nuisance parameters. When the study only has type 1 right censoring, though, the contribution to the likelihood by the censoring distribution will always be 1, so the censoring distribution can be factored out of the likelihood and does not need to be estimated. In order to solve for the nonparametric maximum estimated likelihood survival function estimator $F$ using the estimated likelihood we developed, we first note that the maximum estimate will be a step function that is continuous from the right with left limits that falls only at event times observed in the validation set, $t_1, \cdots, t_{\tilde{K}}$, where $\tilde{K}$ is the number of unique true event times. Similarly, if the censoring distribution is being estimated, the maximum estimator will be a step function that is continuous from the right with left limits that falls only at censoring times observed in the validation set. To solve for the parameters, we used the Nelder-Mead algorithm to conduct constrained maximization. We required that both $F$ and $G$ survival functions are monotonically non-increasing as time increases and are bounded between 0 and 1. In the case where the parameter space is one-dimensional, meaning there is only one observed event time in the validation set data and only type 1 censoring, we used the Brent algorithm. To obtain initial estimates for the event distribution parameters, we used the complete-case Kaplan-Meier estimates based on the true observed times and true event indicators from the vali-

dation set. Initial parameters for the censoring distribution were determined by the complete-case Kaplan-Meier estimates calculated by inverting the event indicator to obtain a censoring indicator. Let $\hat{F}(t_{\tilde{k}})$ represent the event distribution estimates obtained from the algorithm for $\tilde{k} = 1, \cdots, \tilde{K}$. The maximum estimated likelihood survival function estimator is then the step function that takes value 1 in the interval $[0, t_1)$, $\hat{F}(t_{\tilde{k}})$ in each interval $[t_{\tilde{k}}, t_{\tilde{k}+1})$ for $\tilde{k} = 1, \cdots, \tilde{K} - 1$, and $\hat{F}(t_{\tilde{K}})$ in the interval $[t_{\tilde{K}}, x_K]$, where $x_K$ is the last true observed time and may be equal to $t_{\tilde{K}}$ if a true event occurs at the last true observed time. The estimator is considered undefined after $x_K$.

## 2.3. Asymptotic Properties of the Proposed Nonparametric Maximum Estimated Likelihood Estimator

The asymptotic properties of the proposed estimator refer to the situation when the total number of subjects $n \to \infty$. As long as the proportion of subjects in the validation set to the total number of subjects does not have a zero limit, $\lim_{n\to\infty} \frac{n_v}{n} = p_V > 0$, similar arguments as in Theorem 3.1 of Pepe (1992) imply that $\hat{F}(t)$ is a consistent estimator for $F(t)$ for all times $t$. Although $\hat{F}(t)$ is only estimated at observed event times, $t_1, \cdots, t_{\tilde{K}}$, this set of observed event times will approach the set of all possible observed event times, or $\tilde{K} \to K$ as $n \to \infty$. Because we have discrete time points, $F(t)$ is also a step function that can be defined by the survival function values at each time point, $F(t_1), \cdots, F(t_K)$, and we have that

$$\sqrt{n}\left[\begin{pmatrix} \hat{F}(t_1) \\ \hat{F}(t_2) \\ \vdots \\ \hat{F}(t_K) \end{pmatrix} - \begin{pmatrix} F(t_1) \\ F(t_2) \\ \vdots \\ F(t_K) \end{pmatrix}\right]$$

converges to a zero-mean Gaussian random variable in distribution with asymptotic variance covariance matrix equal to $\Sigma_F$, where $\Sigma_F$ is the top left $K \times K$ quadrant of the full variance covariance matrix

$$\Sigma = \mathcal{I}^{-1} + \frac{(1 - p_V)^2}{p_V} \mathcal{I}^{-1} \mathcal{K} \mathcal{I}^{-1}, \tag{2.10}$$

where $\mathcal{I}$ is the information matrix based on the (non-estimated) log likelihood and $\mathcal{K}$ is the expected conditional variance of the non-validation contribution to the log likelihood (Pepe, 1992),

$$\mathcal{K} = \mathrm{E}\left[\mathrm{Var}\left\{ \frac{\partial \log P(X^*, \delta^*)}{\partial \theta} \middle| X, \delta \right\}\right] \tag{2.11}$$

for parameters $\theta = \{F, G\}$. The first term in the $\Sigma$ variance expression represents the variance component based on the maximum likelihood estimator and the second term represents a penalty from estimating the likelihood with empirical probabilities. The $\mathcal{I}$ and $\mathcal{K}$ matrices can be estimated consistently by

$$\hat{\mathcal{I}} = \frac{1}{n} \frac{\partial^2 \log \hat{L}}{\partial \theta^2}\bigg|_{\theta=\hat{\theta}} \tag{2.12}$$

for maximum estimated likelihood estimates $\hat{\theta} = \{\hat{F}, \hat{G}\}$ and

$$\hat{\mathcal{K}} = \frac{1}{n_V} \sum_{i \in V} \hat{Q}_i \hat{Q}_i^T \bigg|_{\theta=\hat{\theta}}, \tag{2.13}$$

where

$$\hat{Q}_i = \frac{1}{n - n_V} \frac{1}{\hat{P}(X_i, \delta_i)} \sum_{j \in \bar{V}} \left[ \left\{ I(X_j^* = X_i^*, \delta_j^* = \delta_i^*) - \hat{P}(X_j^*, \delta_j^* | X_i, \delta_i) \right\} \cdot \left\{ \frac{D(X_i, \delta_i)}{\hat{P}(X_j^*, \delta_j^*)} - \frac{\hat{D}(X_j^*, \delta_j^*)}{\hat{P}^2(X_j^*, \delta_j^*)} P(X_i, \delta_i) \right\} \right] \tag{2.14}$$

and

$$D(X_i, \delta_i) = \frac{\partial P(X_i, \delta_i)}{\partial \theta} \tag{2.15}$$

$$\hat{D}(X_j^*, \delta_j^*) = \sum_{k=1}^{K} \sum_{\delta=0}^{1} \frac{\partial P(X, \delta)}{\partial \theta} \hat{P}(X_j^*, \delta_j^* | X, \delta). \tag{2.16}$$

In practice, derivatives in the variance expression can be calculated numerically. We found that the numerical derivatives were sometimes unable to be computed or led to negative variances with data that had large amounts of missingness or large numbers of parameters to estimate. In these cases, bootstrapped variance estimates can be calculated or analytical forms of the derivatives should be used.

## 2.4. Simulations

Our proposed survival function estimator is motivated by the fact that true endpoints are missing for some subjects while uncertain endpoints are available for all subjects and carry useful information for survival function estimation. In order to assess the performance of our proposed survival function estimator, we conducted a series of simulation studies. We simulated the true event time from a discrete uniform distribution, $T \sim$Unif$[1, 8]$, where survival time can only take integer values, and assumed right censoring at $C = 7$. The uncertain time to event was calculated as $T^* = T + \epsilon$, where $\epsilon \sim$Unif$[0, \zeta]$ and $\epsilon$ is independent of $T$. The maximum integer value of the discrete uniform distribution for $\epsilon$ was calculated as $\zeta = \left\lfloor \sqrt{63 \cdot \frac{1-\rho^2}{\rho^2} + 1} - 1 \right\rfloor$, where $\lfloor a \rfloor$ represents the largest integer not greater than $a$, where $\rho$ represents the correlation between $T$ and $T^*$. The expression for $\zeta$ was computed using the definition of correlation between $T$ and $T^*$, independence of $\epsilon$ and $T$, and variance expressions for $T$ and $\epsilon$. Mathematical details of the derivation can be found in Appendix A.2. We considered correlations of $\rho \in \{0.01, 0.25, 0.50, 0.75, 1\}$. We set the right-censoring time for the uncertain endpoint also at $C^* = 7$. To create a representative validation subsample, we simulated data missing completely at random (MCAR) by randomly selecting a proportion $r \in \{0.25, 0.50, 0.75\}$ of the sample to be missing true endpoints. We used total sample sizes of $n \in \{200, 500\}$ and conducted 500 repetitions of the simulation for each set of parameter values.

For each simulation, we used the proposed method to calculate survival function estimates at each observed time. We also calculated complete-case Kaplan-Meier estimates using only true endpoints in the validation set, the naïve Kaplan-Meier estimates using only uncertain endpoints from all subjects, and the true Kaplan-Meier estimates using true endpoints from all subjects (which would be unavailable in real data). For the proposed estimator, the complete-case Kaplan-Meier estimator, and the naïve Kaplan-Meier estimator, we calculated estimated bias (parameter estimate$-$true parameter values), observed sample standard deviations (SD), estimated standard errors ($\widehat{\text{SE}}$), relative efficiency (RE) compared to the true Kaplan-Meier estimator (where lower RE implies greater efficiency and RE equal to 1 implies optimal efficiency), mean squared error (MSE) estimates, and 95% coverage (Cov) at each of the observed time points. Each statistic was then averaged over all time points. We note that for all simulations presented in Tables 2.1, 2.2, and 2.3, the observed sample standard deviation corresponds well with the standard error estimates from the asymptotic theory for the proposed estimator.

Table 2.1 shows the results from the simulation study with type 1 censoring and $n = 200$. The proposed estimator behaves similarly to the complete-case Kaplan-Meier estimator in terms of bias. Both have little bias, whereas the naïve Kaplan-Meier estimator is heavily biased. When the proportion of missingness is low or moderate ($r = 0.25$ or $r = 0.50$), the relative efficiency of our proposed estimator is similar to that of the complete-case Kaplan-Meier estimator when correlation is low and improves until it reaches optimal efficiency with correlation of 1, which can be interpreted as the situation where the uncertain outcome has no measurement error. The MSE of the proposed estimator is also similar to then becomes smaller than the MSE of the complete-case Kaplan-Meier estimator as correlation increases, and it is consistently smaller than the MSE of the naïve Kaplan-Meier estimator. This demonstrates that using the internal validation subsample can reduce the bias of survival estimates compared to using only uncertain endpoints and that using uncertain endpoints in the non-validation subsample can improve efficiency compared to using only true endpoints. When the amount of missing true outcomes is high ($r = 0.75$), though, our proposed estimator is actually slightly less efficient than the complete-case Kaplan-Meier estimator at low correlations between outcomes.

We saw similar results for simulations with $n = 500$, as shown in Appendix A.4. In addition, we tested the performance of our method at smaller sample sizes, $n \in \{10, 20, 30, \cdots\}$, to determine an approximate threshold for the number of subjects per parameter or events per variable (EPV) needed for estimation. We calculated the EPV as the smallest number of events in the validation set divided by 7, the number of parameters to estimate, such that average bias was less than 0.01 and average coverage was between 93% and 97%. Through these simulation studies, we found an EPV of 4. We also increased the proportion of censored subjects (results not shown) by setting an earlier censoring time for both endpoints and arrived at the same conclusions. Although we assumed only non-negative measurement error of the uncertain endpoint for our simulations to demonstrate the potentially large bias in the naïve estimator and to better control the correlation between outcomes, we also conducted simulations allowing for negative or positive measurement error and the results (not shown) for our estimator and the complete-case estimator are similar.

To compare the efficiency between our proposed estimator and the complete-case Kaplan-Meier estimator over various amounts of missingness, we computed the relative efficiencies (averaged over times) at 5% increments of the percentage of missingness of true endpoints for correlations of

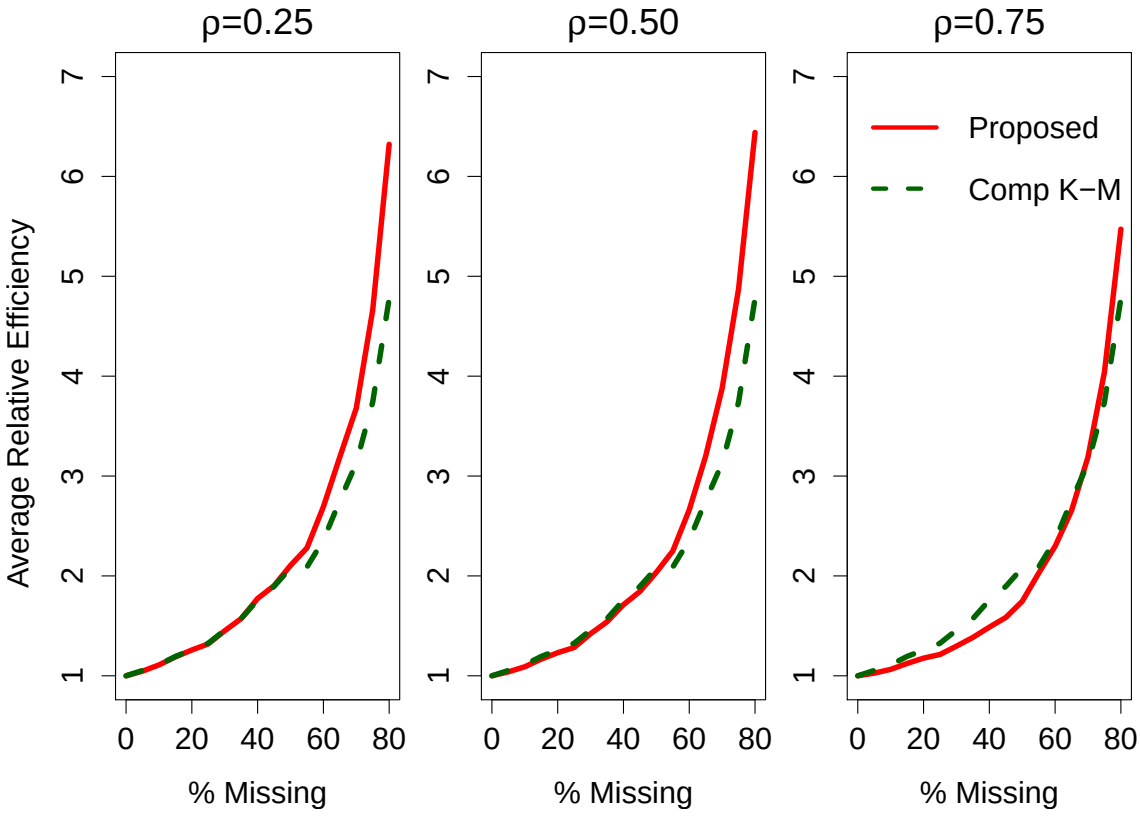
Table 2.1: Simulation Results for Type 1 Censoring and $n = 200$

| $r$ | $\rho$ | Method | Bias $\times 10^{-3}$ | SD | $\hat{SE}$ | MSE $\times 10^{-3}$ | RE | Cov |
|---|---|---|---|---|---|---|---|---|
| | | Proposed | -0.26 | 0.035 | 0.035 | 1.27 | 1.36 | 0.96 |
| | 0.01 | Comp K-M | -0.55 | 0.035 | 0.035 | 1.27 | 1.36 | 0.95 |
| | | Naïve K-M | 498.41 | 0.003 | 0.002 | 310.32 | 0.01 | 0.00 |
| | | Proposed | 0.79 | 0.035 | 0.035 | 1.28 | 1.36 | 0.96 |
| | 0.25 | Comp K-M | -0.55 | 0.035 | 0.035 | 1.27 | 1.36 | 0.95 |
| | | Naïve K-M | 449.43 | 0.014 | 0.014 | 247.32 | 0.28 | 0.00 |
| | | Proposed | 0.57 | 0.035 | 0.034 | 1.26 | 1.34 | 0.96 |
| 25 | 0.50 | Comp K-M | -0.55 | 0.035 | 0.035 | 1.27 | 1.36 | 0.95 |
| | | Naïve K-M | 384.50 | 0.020 | 0.020 | 175.53 | 0.56 | 0.00 |
| | | Proposed | 0.32 | 0.034 | 0.033 | 1.18 | 1.26 | 0.96 |
| | 0.75 | Comp K-M | -0.55 | 0.035 | 0.035 | 1.27 | 1.36 | 0.95 |
| | | Naïve K-M | 285.90 | 0.025 | 0.025 | 91.52 | 0.83 | 0.00 |
| | | Proposed | 0.02 | 0.030 | 0.030 | 0.94 | 1.00 | 0.96 |
| | 1.00 | Comp K-M | -0.55 | 0.035 | 0.035 | 1.27 | 1.36 | 0.95 |
| | | Naïve K-M | 0.03 | 0.030 | 0.030 | 0.94 | 1.00 | 0.95 |
| | | Proposed | -0.15 | 0.043 | 0.043 | 1.88 | 1.99 | 0.95 |
| | 0.01 | Comp K-M | -1.01 | 0.043 | 0.042 | 1.87 | 1.98 | 0.95 |
| | | Naïve K-M | 498.41 | 0.003 | 0.002 | 310.32 | 0.01 | 0.00 |
| | | Proposed | 4.62 | 0.044 | 0.042 | 2.02 | 2.14 | 0.95 |
| | 0.25 | Comp K-M | -1.01 | 0.043 | 0.042 | 1.87 | 1.98 | 0.95 |
| | | Naïve K-M | 449.43 | 0.014 | 0.014 | 247.32 | 0.28 | 0.00 |
| | | Proposed | 3.10 | 0.044 | 0.042 | 1.94 | 2.05 | 0.95 |
| 50 | 0.50 | Comp K-M | -1.01 | 0.043 | 0.042 | 1.87 | 1.98 | 0.95 |
| | | Naïve K-M | 384.50 | 0.020 | 0.020 | 175.53 | 0.56 | 0.00 |
| | | Proposed | 2.32 | 0.042 | 0.040 | 1.76 | 1.88 | 0.95 |
| | 0.75 | Comp K-M | -1.01 | 0.043 | 0.042 | 1.87 | 1.98 | 0.95 |
| | | Naïve K-M | 285.90 | 0.025 | 0.025 | 91.52 | 0.83 | 0.00 |
| | | Proposed | 0.02 | 0.030 | 0.030 | 0.94 | 1.00 | 0.96 |
| | 1.00 | Comp K-M | -1.01 | 0.043 | 0.042 | 1.87 | 1.98 | 0.95 |
| | | Naïve K-M | 0.03 | 0.030 | 0.030 | 0.94 | 1.00 | 0.95 |
| | | Proposed | 3.35 | 0.061 | 0.060 | 3.86 | 4.11 | 0.96 |
| | 0.01 | Comp K-M | 1.86 | 0.061 | 0.060 | 3.83 | 4.07 | 0.96 |
| | | Naïve K-M | 498.41 | 0.003 | 0.002 | 310.32 | 0.01 | 0.00 |
| | | Proposed | 24.12 | 0.065 | 0.065 | 4.39 | 4.65 | 0.98 |
| | 0.25 | Comp K-M | 1.86 | 0.061 | 0.060 | 3.83 | 4.07 | 0.96 |
| | | Naïve K-M | 449.43 | 0.014 | 0.014 | 247.32 | 0.28 | 0.00 |
| | | Proposed | 21.49 | 0.067 | 0.063 | 4.58 | 4.83 | 0.96 |
| 75 | 0.50 | Comp K-M | 1.86 | 0.061 | 0.060 | 3.83 | 4.07 | 0.96 |
| | | Naïve K-M | 384.50 | 0.020 | 0.020 | 175.53 | 0.56 | 0.00 |
| | | Proposed | 9.84 | 0.061 | 0.058 | 3.83 | 4.13 | 0.95 |
| | 0.75 | Comp K-M | 1.86 | 0.061 | 0.060 | 3.83 | 4.07 | 0.96 |
| | | Naïve K-M | 285.90 | 0.025 | 0.025 | 91.52 | 0.83 | 0.00 |
| | | Proposed | -0.22 | 0.031 | 0.030 | 0.97 | 1.03 | 0.96 |
| | 1.00 | Comp K-M | 1.86 | 0.061 | 0.060 | 3.83 | 4.07 | 0.96 |
| | | Naïve K-M | 0.03 | 0.030 | 0.030 | 0.94 | 1.00 | 0.95 |

$r$ is the percent missing and $\rho$ is the correlation between true and uncertain outcomes. Proposed refers to the proposed estimator, Comp K-M refers to the complete-case Kaplan-Meier estimator, and Naïve K-M refers to the naïve Kaplan-Meier estimator. SD is standard deviation of estimates across simulations, $\hat{SE}$ is estimated standard error of the estimate, MSE is mean squared error, RE is relative efficiency, Cov is 95% coverage, all averaged across time.

$\rho \in \{0.25, 0.50, 0.75\}$ (Figure 2.1). For these simulations, we used a larger sample size of $n = 500$ to ensure that the EPV was adequate even at the largest amounts of missingness. For correlations of 0.50 and 0.75, our proposed estimator is more efficient (lower RE) than the complete-case Kaplan-Meier estimator when the proportion of missing data is low, then the efficiency curves cross and our proposed estimator becomes less efficient. The point of crossing is at a higher percentage of missingness with higher values of the correlation between outcomes. Even with low correlation ($\rho = 0.25$) between outcomes, though, our estimator has similar or lower efficiency than the complete-case Kaplan-Meier estimator when the amount of missingness is 50% or less. This is consistent with Pepe's recommendation for non-survival data with one parameter of interest (Pepe, 1992).

Figure 2.1: Relative Efficiencies by Correlation Between True and Uncertain Endpoints ($\rho$) and Amount of Missingness of True Endpoints




Proposed refers to the proposed estimator and Comp K-M refers to the complete-case Kaplan-Meier estimator. This figure appears in color in the electronic version of this article.

We explored the behavior of our proposed estimator under random censorship by simulating true event times $T \sim$Unif$[1, 8]$, uncertain event times $T^* = T + \epsilon$ where $\epsilon \sim$Unif$[0, 2]$, true censor-

ing times $C \sim$Unif$[5, 7]$, and uncertain censoring times $C^* = C + \gamma$ where $\gamma \sim$Unif$[0, 2]$. These simulations resulted in a small amount of censoring (approximately 30%). We also increased the amount of censoring by simulating true censoring times $C \sim$Unif$[3, 7]$, which resulted in a larger amount of censoring (approximately 50%). The results of these random censoring simulations are shown in Table 2.2. Similar to the results from type 1 censoring, our proposed estimator has little bias compared to the naïve Kaplan-Meier estimator and is more efficient than the complete-case Kaplan-Meier estimator for both small and large amounts of censoring. We saw similar results with $n = 500$ as seen in Appendix A.5.

Table 2.2: Simulation Results for Random Censoring and $n = 200$

| $r$ | $C$ | Method | Bias $\times 10^{-3}$ | SD | $\widehat{\text{SE}}$ | MSE $\times 10^{-3}$ | RE | Cov |
|---|---|---|---|---|---|---|---|---|
| 25 | S | Proposed | 0.24 | 0.035 | 0.034 | 1.25 | 1.17 | 0.96 |
| | | Comp K-M | 0.39 | 0.038 | 0.037 | 1.47 | 1.39 | 0.95 |
| | | Naïve K-M | 119.63 | 0.030 | 0.029 | 15.47 | 0.83 | 0.03 |
| | L | Proposed | 0.57 | 0.040 | 0.038 | 1.65 | 1.20 | 0.96 |
| | | Comp K-M | 0.37 | 0.043 | 0.041 | 1.94 | 1.44 | 0.94 |
| | | Naïve K-M | 119.98 | 0.032 | 0.032 | 15.75 | 0.78 | 0.05 |
| 50 | S | Proposed | -2.18 | 0.040 | 0.041 | 1.64 | 1.54 | 0.95 |
| | | Comp K-M | -0.12 | 0.047 | 0.046 | 2.23 | 2.10 | 0.95 |
| | | Naïve K-M | 119.63 | 0.03 | 0.029 | 15.47 | 0.83 | 0.03 |
| | L | Proposed | -1.99 | 0.049 | 0.044 | 2.68 | 1.84 | 0.96 |
| | | Comp K-M | -0.52 | 0.053 | 0.050 | 2.95 | 2.18 | 0.93 |
| | | Naïve K-M | 119.98 | 0.032 | 0.032 | 15.75 | 0.78 | 0.05 |

$r$ is the percent missing and $C$ is the amount of censoring, where S means small (30%) and L means large (50%). Proposed refers to the proposed estimator, Comp K-M refers to the complete-case Kaplan-Meier estimator, and Naïve K-M refers to the naïve Kaplan-Meier estimator. SD is standard deviation of estimates across simulations, $\widehat{\text{SE}}$ is estimated standard error of the estimate, MSE is mean squared error, RE is relative efficiency, Cov is 95% coverage, all averaged across time.

To test the robustness of the MCAR assumption of the proposed method, we relaxed this assumption and simulated data missing at random (MAR). We defined a missingness indicator $R$, where $R = 1$ denotes a missing true endpoint and $R = 0$ denotes a non-missing true endpoint, based on

the uncertain indicator $\delta^*$ such that

$$R|(\delta^* = 0) = \begin{cases} 1 & \text{with probability } p_R \\ 0 & \text{with probability } 1 - p_R \end{cases}$$

$$R|(\delta^* = 1) = \begin{cases} 1 & \text{with probability } 1 - p_R \\ 0 & \text{with probability } p_R \end{cases}$$

for probability $p_R = 0.60$. This implies that the probability of missingness of the true endpoint depends on the observed censoring indicator of the uncertain endpoint. In the AD example, this would imply that subjects who are clinically determined to be non-AD during the study are more likely to miss the CSF biomarker endpoint. In the results from the MAR data in Table 2.3, we see that both the proposed estimator and the complete-case Kaplan-Meier estimator is sometimes slightly biased. However, the proposed estimator is less biased than the complete-case Kaplan-Meier estimator, particularly when the correlation between outcomes is very high. Because of these differences in bias, the coverage of the proposed estimator is better than the coverage of the complete-case Kaplan-Meier estimator when the correlation between outcomes is greater than 0.01. We saw similar results with $n = 500$ as seen in Appendix A.6.

## 2.5. Application to the Alzheimer's Disease Neuroimaging Initiative Study

We illustrated our method by considering data (retrieved on July 26, 2013) from the ongoing ADNI study (Weiner et al., 2012). See Appendix A.3 for more detailed information about the ADNI study. Participants in this study were seen every 6 months until the end of two years, then annually thereafter, at which time clinical diagnoses of non-AD (cognitively normal or MCI) or AD were assessed. These follow-up times were predetermined by study design, and thus discrete survival estimates would be appropriate in this study. The current study includes data from participants in the ADNI-1 and ADNI-GO segments of the ADNI study. For those who agreed to a lumbar puncture, CSF assays were performed and $A\beta$ protein concentrations were measured. Participants with an $A\beta$ biomarker value greater than 192 pg/ml were classified as non-AD at baseline and those with an $A\beta$ value less than or equal to 192 pg/ml were classified as AD at baseline (Shaw et al., 2009). There were 186 patients who were non-AD at the time of enrollment according to both the clinical diagnosis and CSF diagnosis. For each patient, the time to clinical AD or last follow-up was

Table 2.3: Simulation Results for Data Missing at Random and $n = 200$

| Censoring | $\rho/C$ | Method | Bias $\times 10^{-3}$ | SD | $\hat{SE}$ | MSE $\times 10^{-3}$ | RE | Cov |
|---|---|---|---|---|---|---|---|---|
| Type 1 | 0.01 | Proposed | 2.53 | 0.048 | 0.048 | 2.31 | 2.47 | 0.96 |
| | | Comp K-M | 0.91 | 0.048 | 0.048 | 2.31 | 2.47 | 0.95 |
| | | Naïve K-M | 498.41 | 0.003 | 0.002 | 310.32 | 0.01 | 0.00 |
| | 0.25 | Proposed | 17.06 | 0.048 | 0.049 | 2.38 | 2.56 | 0.97 |
| | | Comp K-M | -11.75 | 0.046 | 0.046 | 2.17 | 2.31 | 0.94 |
| | | Naïve K-M | 449.43 | 0.014 | 0.014 | 247.32 | 0.28 | 0.00 |
| | 0.50 | Proposed | 21.47 | 0.047 | 0.046 | 2.22 | 2.38 | 0.98 |
| | | Comp K-M | -26.55 | 0.045 | 0.045 | 2.06 | 2.19 | 0.90 |
| | | Naïve K-M | 384.50 | 0.020 | 0.020 | 175.53 | 0.56 | 0.00 |
| | 0.75 | Proposed | 16.27 | 0.042 | 0.041 | 1.78 | 1.94 | 0.96 |
| | | Comp K-M | -44.30 | 0.043 | 0.042 | 1.89 | 2.00 | 0.81 |
| | | Naïve K-M | 285.90 | 0.025 | 0.025 | 91.52 | 0.83 | 0.00 |
| | 1.00 | Proposed | 0.02 | 0.030 | 0.030 | 0.94 | 1.00 | 0.96 |
| | | Comp K-M | -22.56 | 0.039 | 0.039 | 1.57 | 1.65 | 0.88 |
| | | Naïve K-M | 0.03 | 0.030 | 0.030 | 0.94 | 1.00 | 0.95 |
| Random | S | Proposed | 0.32 | 0.039 | 0.044 | 1.62 | 1.49 | 0.96 |
| | | Comp K-M | -33.92 | 0.043 | 0.042 | 1.87 | 1.78 | 0.86 |
| | | Naïve K-M | 119.63 | 0.03 | 0.029 | 15.47 | 0.83 | 0.03 |
| | L | Proposed | 1.66 | 0.047 | 0.044 | 2.49 | 1.69 | 0.96 |
| | | Comp K-M | -41.74 | 0.048 | 0.047 | 2.31 | 1.81 | 0.83 |
| | | Naïve K-M | 119.98 | 0.032 | 0.032 | 15.75 | 0.78 | 0.05 |

Censoring is the type of the censoring mechanism and $\rho/C$ either represents the correlation $\rho$ between true and uncertain outcomes or represents the amount of censoring, where S means small (30%) and L means large (50%). Proposed refers to the proposed estimator, Comp K-M refers to the complete-case Kaplan-Meier estimator, and Naïve K-M refers to the naïve Kaplan-Meier estimator. SD is standard deviation of estimates across simulations, $\hat{SE}$ is estimated standard error of the estimate, MSE is mean squared error, RE is relative efficiency, Cov is 95% coverage, all averaged across time.

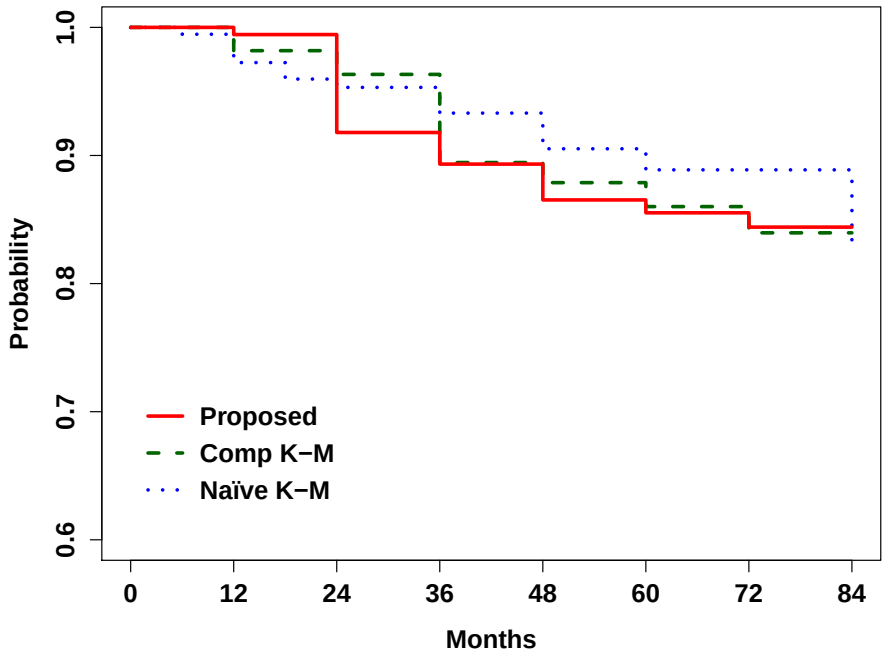
recorded to obtain an uncertain, mismeasured outcome on all patients. A subset of 110 patients continued to have CSF assays performed annually. For these 110 patients in the validation set, patients were classified as non-AD or AD at each time point using the same cutoff of 192 pg/ml and the true time to AD or last follow-up was also recorded. Thus, patients with any CSF assays during follow-up were considered to be in the validation set and those with no CSF assays during follow-up were considered to be in the non-validation set, or $n_V = 110$ and $n = 186$ using the notation of Section 2.2.

First, we assessed the missingness mechanism in the data. We used a log-rank test to compare the survival functions for time to clinical AD diagnosis between the non-validation set and the validation set. The $\chi^2$ test statistic was 0.2 with 1 degree of freedom, yielding a p-value of 0.662. We also used Fisher's exact test to test for an association between the clinical event indicator and missingness. The p-value was 1. Further, because we used all available longitudinal CSF assays, those who were missing CSF diagnoses were missing immediately after baseline. Since all subjects begin as non-AD at baseline, the missingness could not be dependent on baseline CSF or clinical diagnoses. Therefore, we did not find strong evidence against the MCAR assumption.

Figure 2.2 shows the estimated survival functions using our proposed estimator which maximized the estimated likelihood, the complete-case Kaplan-Meier estimator which only uses 110 CSF diagnoses, and the naïve Kaplan-Meier estimator which only uses the 186 clinical diagnoses. The three survival functions are very similar until 36 months, at which time the naïve Kaplan-Meier estimate begins to diverge from the other two survival curves. With higher survival probabilities, the naïve estimate overestimates the probability of being AD-free after 36 months compared to the proposed estimator and complete-case Kaplan-Meier estimator. Since the naïve estimate is based on only clinical diagnoses, this would indicate that abnormality of $A\beta$ occurred earlier than cognitive impairment. This finding is consistent with the recent theoretical model of AD pathology developed by Jack et al. (2010).

Table 2.4 shows the standard error estimates at each time point. The standard errors of the proposed estimate are similar to or smaller than those of the complete-case Kaplan-Meier estimate at all time points. This further supports the conclusion that the proposed estimator helps to improve efficiency relative to the complete-case estimator.

Figure 2.2: Data Example Survival Function Estimates for Time to AD




Proposed refers to the proposed estimator, Comp K-M refers to the complete-case Kaplan-Meier estimator, and Naïve K-M refers to the naïve Kaplan-Meier estimator. This figure appears in color in the electronic version of this article.

## 2.6. Discussion

We proposed a nonparametric maximum likelihood estimator for the discrete survival function in the presence of uncertain endpoints by using an internal validation subsample. We allowed for random censoring for survival outcomes by incorporating a censoring distribution in the likelihood, showed that the survival function estimator is a step function that drops only at observed event times, and proved that the proposed estimator is consistent and asymptotically normal at each discrete time point. We evaluated the finite sample performance of the proposed estimator through extensive simulations. We found that the proposed estimator has little bias and can improve efficiency relative to the complete-case Kaplan-Meier estimator. It can also reduce bias compared to the naïve Kaplan-Meier estimator. The proposed estimator also works better than the complete-case and naïve estimators under departure from the MCAR assumption.

The efficiency gains of the proposed estimator have useful implications in clinical trials. A true

Table 2.4: Data Example Standard Error Estimates

| Month | Proposed Estimator | Complete-Case Kaplan-Meier | Naïve Kaplan-Meier |
|---|---|---|---|
| 6 | 0.000 | 0.000 | 0.005 |
| 12 | 0.008 | 0.013 | 0.012 |
| 18 | 0.008 | 0.013 | 0.016 |
| 24 | 0.022 | 0.019 | 0.017 |
| 36 | 0.036 | 0.036 | 0.023 |
| 48 | 0.038 | 0.040 | 0.031 |
| 60 | 0.040 | 0.045 | 0.036 |
| 72 | 0.046 | 0.051 | 0.036 |
| 84 | 0.046 | 0.051 | 0.074 |

outcome may be costly to obtain on all subjects, but using the proposed method can incorporate a less costly uncertain outcome assessed on all subjects and the true outcomes on a smaller subsample. Compared to obtaining true outcomes on all subjects which can be very costly or using a complete-case estimator on the smaller subsample, our estimator can reduce costs of the trial without sacrificing power.

The proposed approach does not require that only subjects with positive uncertain endpoints (e.g., having clinical AD in our data example) can be validated in contrast to previous literature. Our approach allows that all subjects can have the opportunity to be validated. Through simulations, we found that the efficiency gains of our proposed estimator depends on both the correlation between the uncertain and true outcome and the size of the validation sample. However, in general, the proposed estimator seems to work well when the size of the validation sample is 50% or more of the total sample size. The proposed method can be used with data that have both type 1 right censoring and random right censoring, whereas previous methods only allowed type 1 right censoring. The proposed method also assumes that study subjects are seen at predetermined time points and relies on a discrete time framework. In studies where subjects are evaluated at any time, the proposed estimator may not improve efficiency compared to the complete-case Kaplan-Meier estimator. For this situation, a modified approach must be developed.

The proposed method only estimates a single survival function. A natural extension of the method would be a semiparametric version that is able to incorporate covariates and conduct between-group comparisons. The extension of our proposed method for a proportional hazards model with a binary or continuous covariate of interest is discussed in Chapter 3.

As early detection of Alzheimer's disease and other chronic diseases becomes increasingly important, but event outcomes may be hard to obtain for everyone, we recommend collecting an internal validation sample when the measures of the event outcome are uncertain so that statistical analysis can be improved with greater accuracy and power.

# CHAPTER 3

# SEMIPARAMETRIC SURVIVAL ANALYSIS WITH UNCERTAIN ENDPOINTS USING AN INTERNAL VALIDATION SUBSAMPLE

## 3.1. Introduction

In epidemiological studies and clinical trials, interest often lies in comparing the effects of treatment on time to an event. The Cox proportional hazards model is a common, standard method of survival analysis for analyzing true outcome data, but true outcomes are often unavailable due to invasiveness or cost restrictions. For example, in the pathological diagnosis of Alzheimer's disease (AD), the outcome of interest may be a cerebral spinal fluid (CSF) diagnosis. However, the CSF assay requires a lumbar puncture to measure A$\beta$ protein concentrations in the spinal fluid, which is considered too painful for some patients. In these cases, true outcomes data may be supplemented by alternative outcomes that measure the true outcomes with some error. In the case of diagnosing AD, a clinical diagnosis based on cognitive tests may be used. The clinical diagnosis presents differently from the CSF diagnosis and therefore measures the true outcome with error because clinical symptoms of AD are easily mistaken for other types of dementia. However, clinical diagnoses are more widely available than the CSF diagnoses. Using both the uncertain, mismeasured outcome on all subjects and the true outcome on a subsample of subjects, called the validation sample, estimates of covariate effects can be improved.

Previous methods for estimating survival outcomes in these situations assumed known mismeasurement rates of the uncertain outcome, allowed only positive uncertain outcomes to be validated with an assessment of the true outcome, and/or only allowed for fixed censoring (Richardson and Hughes, 2000; Meier et al., 2003; Balasubramanian and Lagakos, 2001; Magaret, 2008). For example, Magaret (2008) adapted a method first introduced by Pepe (1992) for discrete survival data and only discrete covariates using the proportional hazards model. Pepe's method involved an estimated likelihood method that incorporates information from both uncertain outcomes and true outcomes, where estimation is performed for the conditional probability of the uncertain outcome given the true outcome (Pepe, 1992). Zee and Xie (Chapter 2) adapted the estimated likelihood method to estimate a survival function allowing any subject to be validated and allowing for either

fixed or random censoring mechanisms.

In this paper, we extend the work of Zee and Xie (Chapter 2, under revision for *Biometrics*) to a semiparametric estimated likelihood method to estimate a parameter representing a covariate effect for data with uncertain outcomes on all subjects and true outcomes on a subset. We assume a proportional hazards model and estimate the log hazard ratio of the survival outcome comparing different covariate values. Unlike Magaret's (2008) method which only considered discrete covariates of interest, we consider both a binary or a continuous covariate. Although we express our approach with a binary categorical variable for ease of notation, the method can be easily modified to consider categorical variables with more than two levels. For the continuous covariate, we use a smooth kernel type estimator within the likelihood. The proposed estimator of the log hazard ratio is consistent and asymptotically normal. The rest of the article is organized as follows. Section 3.2 describes the estimated likelihood and asymptotic properties for a binary covariate. Section 3.3 describes the method and asymptotic properties for a continuous covariate. Section 3.4 contains results of our simulation study. In Section 3.5, we demonstrate the use of our method using data from the Alzheimer's Disease Neuroimaging Initiative to estimate covariate effects. Finally, we summarize our results and discuss implications of using our proposed method in Section 3.6.

## 3.2. Semiparametric Estimated Likelihood with a Binary Covariate

### *3.2.1. Maximum Estimated Likelihood Estimation*

Let $T$ represent the true time to event and $C$ represent the true right censoring time, with event indicator $\delta = I(T \leq C)$. Similarly, let $T^*$ represent the uncertain time to event and $C^*$ be the uncertain right censoring time, with indicator $\delta^* = I(T^* \leq C^*)$. Define $X = \min\{T, C\}$ and $X^* = \min\{T^*, C^*\}$. Then $X$ and $X^*$ represent the true and uncertain observed times, respectively. Let $X_k$ represent the $k$th unique, ordered observed true time point for $k = 1, \cdots, K$, where $K$ is the total number of unique true observed times. Let $F_0$ represent the baseline survival function of the true time to event. We assume a proportional hazards model with $F(t) = F_0(t)^{\exp(\beta Z)}$ where $Z \in \{0, 1\}$ is the binary covariate of interest and $\beta$ represents the log hazard ratio of the event comparing $Z = 1$ to $Z = 0$. We assume that the covariate is available for all subjects. Finally, we assume independent censoring, and to allow for random censoring, we let $G$ represent the censoring survival function.

Let $V$ represent the validation set, where both the uncertain and true outcomes are available. There are $n_V$ subjects in the validation set. Then $\bar{V}$ is the non-validation set, where only the uncertain outcome is available and the true outcome is missing. The estimated likelihood would then be

$$\hat{L}(\beta, F_0, G) \propto \prod_{i \in V} P(X_i, \delta_i | Z_i) \prod_{j \in \bar{V}} \hat{P}(X_j^*, \delta_j^* | Z_j). \tag{3.1}$$

where

$$\hat{P}(X_j^*, \delta_j^* | Z_j) = \sum_{k=1}^{K} \sum_{\delta=0}^{1} P(x_k, \delta | Z_j) \hat{P}(X_j^*, \delta_j^* | x_k, \delta, Z_j). \tag{3.2}$$

The conditional probability is estimated empirically by

$$\hat{P}(X_j^*, \delta_j^* | x_k, \delta, Z_j) = \frac{\hat{P}(X_j^*, \delta_j^*, x_k, \delta, Z_j)}{\hat{P}(x_k, \delta, Z_j)} \tag{3.3}$$

$$= \frac{\frac{1}{n_V} \sum_{i \in V} I(X_i^* = X_j^*, \delta_i^* = \delta_j^*, X_i = x_k, \delta_i = \delta, Z_i = Z_j)}{\frac{1}{n_V} \sum_{i \in V} I(X_i = x_k, \delta_i = \delta, Z_i = Z_j)} \tag{3.4}$$

where $I(\cdot)$ is the indicator function. In the estimated likelihood function, only the true outcome contributes to the likelihood for those in the validation set, implying that uncertain outcomes do not provide any additional information when the true outcome is known. However, the uncertain outcomes for those in the validation set are still used to estimate the relationship between the uncertain and true outcomes, which are then used to weight likelihood contributions for those in the non-validation set.

Then for subjects $i \in V$, the contribution to the likelihood is

$$P(X_i, \delta_i | Z_i) = \left\{ F_0(x_{k_i - 1})^{\exp(\beta Z_i)} - F_0(x_{k_i})^{\exp(\beta Z_i)} \right\}^{\delta_i} \left\{ F_0(x_{k_i})^{\exp(\beta Z_i)} \right\}^{1 - \delta_i}$$

$$\cdot G(x_{k_i - 1})^{\delta_i} \left\{ G(x_{k_i - 1}) - G(x_{k_i}) \right\}^{1 - \delta_i} \tag{3.5}$$

$$\propto \left\{ F_0(x_{k_i - 1})^{\exp(\beta Z_i)} - F_0(x_{k_i})^{\exp(\beta Z_i)} \right\}^{\delta_i} \left\{ F_0(x_{k_i})^{\exp(\beta Z_i)} \right\}^{1 - \delta_i} \tag{3.6}$$

where $x_{k_i}$ is the observed time for subject $i$. For subjects $j \in \bar{V}$, the contribution to the likelihood is

$$\hat{P}(X_j^*, \delta_j^* | Z_j) = \sum_{k=1}^{K} \sum_{\delta=0}^{1} \left[ \left\{ F_0(x_{k-1})^{\exp(\beta Z_j)} - F_0(x_k)^{\exp(\beta Z_j)} \right\}^{\delta} \left\{ F_0(x_k)^{\exp(\beta Z_j)} \right\}^{1-\delta} \right.$$
$$\cdot G(x_{k-1})^{\delta} \{ G(x_{k-1}) - G(x_k) \}^{1-\delta}$$
$$\left. \cdot \frac{\frac{1}{n_V} \sum_{i \in V} I(X_i^* = X_j^*, \delta_i^* = \delta_j^*, X_i = x_k, \delta_i = \delta, Z_i = Z_j)}{\frac{1}{n_V} \sum_{i \in V} I(X_i = x_k, \delta_i = \delta, Z_i = Z_j)} \right]. \quad (3.7)$$

The estimated likelihood is a function of the log hazard ratio, $\beta$, and possible survival function values for the baseline event distribution and censoring distribution at each time point. We maximize the estimated likelihood jointly over all possible parameter values. As in the nonparametric case, the maximum estimated likelihood estimate for the event (censoring) survival function is a step function that falls only at event (censoring) times observed in the validation set. We solve for parameters using the Nelder-Mead algorithm with constraints on both $F_0$ and $G$ survival functions to be monotonically non-increasing as time increases and bounded between 0 and 1. To obtain initial estimates for the event distribution parameters, we used the complete-case Kaplan-Meier estimates based on the true observed times and true event indicators from the validation set. Initial parameters for the censoring distribution were determined by the complete-case Kaplan-Meier estimates calculated by inverting the event indicator to obtain a censoring indicator. The initial parameter for the covariate effect is set at 0 and is unconstrained.

### *3.2.2. Asymptotic Properties of $\hat{\beta}$*

The asymptotic properties of the proposed estimator refer to the situation when the total number of subjects $n \to \infty$. As long as the proportion of subjects in the validation set to the total number of subjects does not have a zero limit, $\lim_{n\to\infty} \frac{n_V}{n} = p_V > 0$, similar arguments as in Theorem 3.1 of Pepe (1992) imply that $\hat{\beta}$ is a consistent estimator for $\beta$ and

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{D} N(0, \sigma^2)$$

where $\sigma^2$ is the [1,1] element of the full variance covariance matrix

$$\Sigma = \mathcal{I}^{-1} + \frac{(1 - p_V)^2}{p_V} \mathcal{I}^{-1} \mathcal{K} \mathcal{I}^{-1}, \quad (3.8)$$

where $\mathcal{I}$ is the information matrix based on the (non-estimated) log likelihood and $\mathcal{K}$ is the expected conditional variance of the non-validation contribution to the log likelihood (Pepe, 1992),

$$\mathcal{K} = \mathrm{E}\left[\mathrm{Var}\left\{\left.\frac{\partial \log P(X^*, \delta^*|Z)}{\partial\theta}\right| X, \delta, Z\right\}\right] \tag{3.9}$$

for parameters $\theta = \{\beta, F, G\}$. The first term in the $\Sigma$ variance expression represents the variance component based on the maximum likelihood estimator and the second term represents a penalty from estimating the likelihood with empirical probabilities. The $\mathcal{I}$ and $\mathcal{K}$ matrices can be estimated consistently by

$$\hat{\mathcal{I}} = \frac{1}{n}\left.\frac{\partial^2 \log \hat{L}}{\partial\theta^2}\right|_{\theta=\hat{\theta}} \tag{3.10}$$

for maximum estimated likelihood estimates $\hat{\theta} = \{\hat{\beta}, \hat{F}, \hat{G}\}$ and

$$\hat{\mathcal{K}} = \left.\frac{1}{n_V}\sum_{i\in V}\hat{Q}_i\hat{Q}_i^T\right|_{\theta=\hat{\theta}}, \tag{3.11}$$

where

$$\hat{Q}_i = \frac{1}{n-n_V}\frac{1}{\hat{P}(X_i,\delta_i,Z_i)}\sum_{j\in\bar{V}}\left[\left\{I(X_j^*=X_i^*,\delta_j^*=\delta_i^*) - \hat{P}(X_j^*,\delta_j^*|X_i,\delta_i,Z_i)\right\}I(Z_i=Z_j)\right.$$
$$\left.\cdot\left\{\frac{D(X_i,\delta_i|Z_j)}{\hat{P}(X_j^*,\delta_j^*|Z_j)} - \frac{\hat{D}(X_j^*,\delta_j^*|Z_j)}{\hat{P}^2(X_j^*,\delta_j^*|Z_j)}P(X_i,\delta_i|Z_j)\right\}\right] \tag{3.12}$$

and

$$\hat{P}(X_i,\delta_i,Z_i) = \frac{1}{n_V}\sum_{a\in V} I(X_a=X_i,\delta_a=\delta_i,Z_a=Z_i) \tag{3.13}$$

$$\hat{P}(X_j^*,\delta_j^*|X_i,\delta_i,Z_i) = \frac{\frac{1}{n_V}\sum_{a\in V} I(X_a^*=X_j^*,\delta_a^*=\delta_j^*,X_a=X_i,\delta_a=\delta_i,Z_a=Z_i)}{\frac{1}{n_V}\sum_{a\in V} I(X_a=X_i,\delta_a=\delta_i,Z_a=Z_i)} \tag{3.14}$$

$$D(X_i,\delta_i|Z_j) = \frac{\partial P(X_i,\delta_i|Z_j)}{\partial\theta} \tag{3.15}$$

$$\hat{D}(X_j^*,\delta_j^*|Z_j) = \sum_{k=1}^{K}\sum_{\delta=0}^{1}\frac{\partial P(x_k,\delta|Z_j)}{\partial\theta}\hat{P}(X_j^*,\delta_j^*|x_k,\delta,Z_j). \tag{3.16}$$

In practice, derivatives in the variance expression can be calculated numerically. As in the non-parametric case, we found that the numerical derivatives were sometimes unable to be computed

or led to negative variances with data that had large amounts of missingness or large numbers of parameters to estimate. In these cases, bootstrapped variance estimates can be calculated or analytical forms of the derivatives should be used.

## 3.3. Semiparametric Estimated Likelihood with a Continuous Covariate

### *3.3.1. Maximum Estimated Likelihood Estimation*

For a continuous covariate, using the method above will lead to a 0 contribution by most if not all subjects in the non-validation set. This occurs because the indicator functions used to estimate the conditional probability will be non-zero only if the non-validation set subject's covariate value matches that of some validation set subject. With a continuous covariate, this is extremely unlikely in real data. To address this issue, we use a smooth kernel function to give non-zero values to the probability estimates. The kernel function gives larger values as the non-validation set subject's covariate value is closer to that of validation set subject's covariate values and smaller values when it is further away. Here, we let $Z$ represent the covariate of interest again, but $Z$ is now a continuous random variable. Then $\beta$ represents the log hazard ratio of an event for a one unit change in $Z$.

Let $\phi$ represent a symmetric density function and $h$ represent a pre-specified bandwidth. The estimated likelihood is similar as above, but with conditional probability in the non-validation portion estimated by

$$\hat{P}(X_j^*, \delta_j^* | x_k, \delta, Z_j) = \frac{\frac{1}{n_V} \sum_{i \in V} I(X_i^* = X_j^*, \delta_i^* = \delta_j^*, X_i = x_k, \delta_i = \delta) \frac{1}{h} \phi\left(\frac{Z_i - Z_j}{h}\right)}{\frac{1}{n_V} \sum_{i \in V} I(X_i = x_k, \delta_i = \delta) \frac{1}{h} \phi\left(\frac{Z_i - Z_j}{h}\right)}. \tag{3.17}$$

As in other applications of kernel smoothing techniques, the choice of kernel function and bandwidth may be chosen based on shape of the function, minimization of some measure of mean-squared error, and a desired amount of smoothness (Klein and Moeschberger, 2003; Simonoff, 1996). Common choices of kernel functions include Epanechnikov, Gaussian, Biweight, Triweight, and Uniform, all of which are second-order kernels, meaning the second moment is the first non-zero moment.

### *3.3.2. Asymptotic Properties of $\hat{\beta}$*

The $p$th order kernel function and bandwidth chosen must satisfy the limits $nh^2 \longrightarrow \infty$ and $nh^{2p} \longrightarrow 0$ as $n \longrightarrow \infty$, where the order of the kernel function is the first non-zero moment. Then using

similar arguments as in Pepe (1992), the same asymptotic properties as for the binary covariate hold.

The form for the variance estimator is similar as above, but the $\hat{Q}_i$ portion is instead given by

$$\hat{Q}_i = \frac{1}{n - n_V} \frac{1}{\hat{P}(X_i, \delta_i, Z_i)} \sum_{j \in \bar{V}} \left[ \left\{ I(X_j^* = X_i^*, \delta_j^* = \delta_i^*) - \hat{P}(X_j^*, \delta_j^* | X_i, \delta_i, Z_i) \right\} \frac{1}{h} \phi \left( \frac{Z_i - Z_j}{h} \right) \cdot \left\{ \frac{D(X_i, \delta_i | Z_j)}{\hat{P}(X_j^*, \delta_j^* | Z_j)} - \frac{\hat{D}(X_j^*, \delta_j^* | Z_j)}{\hat{P}^2(X_j^*, \delta_j^* | Z_j)} P(X_i, \delta_i | Z_j) \right\} \right] \tag{3.18}$$

where

$$\hat{P}(X_i, \delta_i, Z_i) = \frac{1}{n_V} \sum_{a \in V} I(X_a = X_i, \delta_a = \delta_i) \frac{1}{h} \phi \left( \frac{Z_a - Z_i}{h} \right) \tag{3.19}$$

$$\hat{P}(X_j^*, \delta_j^* | X_i, \delta_i, Z_i) = \frac{\frac{1}{n_V} \sum_{a \in V} I(X_a^* = X_j^*, \delta_a^* = \delta_j^*, X_a = X_i, \delta_a = \delta_i) \frac{1}{h} \phi \left( \frac{Z_a - Z_i}{h} \right)}{\frac{1}{n_V} \sum_{a \in V} I(X_a = X_i, \delta_a = \delta_i) \frac{1}{h} \phi \left( \frac{Z_a - Z_i}{h} \right)}. \tag{3.20}$$

The inclusion of kernel functions allows probability estimates to be non-zero, as desired. However, there are rare cases where for some non-validation subject $j$, the probability estimate $\hat{P}(X_j^* \delta_j^* | Z_j)$ in the denominators of the $\hat{Q}_i$ expression is extremely close to zero. Although the numerically computed derivatives in the numerators are also extremely close to zero, the contribution to $\hat{Q}_i$ by subject $j$ is extremely large, which makes the variance estimate extremely large. For these cases, it may be necessary to define a zero tolerance such that subjects with probability estimates smaller than that tolerance contribute 0 to the $\hat{Q}_i$ expression.

## 3.4. Simulation Study

To test the performance of our proposed method in estimating a covariate effect, we conducted a series of simulation studies. For a binary covariate, we randomly sampled values $Z \sim$Bernoulli$(0.5)$ and for the continuous case, $Z \sim$N$(0, 1)$. We set the log hazard ratio at $\beta = 1$. We sampled true event times assuming a proportional hazards model with baseline distribution, $T \sim$Unif$[1, 5]$, where survival time can only take integer values. We assumed right censoring at $C = 4$. The uncertain time to event was calculated as $T^* = T + \epsilon$, where $\epsilon \sim$Unif$[0, \zeta]$ and $\epsilon$ is independent of $T$. The maximum integer value of the discrete uniform distribution for $\epsilon$ was calculated as $\zeta = \left\lfloor \sqrt{\text{Var}(T) \cdot \frac{1-\rho^2}{\rho^2} + 1} - 1 \right\rfloor$, where $\lfloor a \rfloor$ represents the largest integer not greater than $a$ and $\rho$

represents the correlation between $T$ and $T^*$. The expression for $\zeta$ was computed using the definition of correlation between $T$ and $T^*$, independence of $\epsilon$ and $T$, and variance expressions for $T$ and $\epsilon$. We considered correlations of $\rho \in \{0.01, 0.25, 0.50, 0.75, 1\}$. We set the right-censoring time for the uncertain endpoint also at $C^* = 4$. To create a representative validation subsample, we simulated data missing completely at random (MCAR). We randomly selected a proportion $r \in \{0.25, 0.50\}$ of the sample to be missing true endpoints, since our previous work suggested efficiency gains with missingness of 50% or less.

For the smooth kernel estimate in the continuous covariate case, we used a standard normal distribution. Silverman (1992) suggested an optimal bandwidth for the standard normal distribution of $h = 0.9\frac{A}{n_V^{1/5}}$ where $A = \min\{\text{standard deviation, interquartile range/1.34}\}$. Because this bandwidth does not satisfy the $nh^2 \longrightarrow \infty$ assumption for the asymptotic properties of the estimator, we used a similar bandwidth of $h = 0.9\frac{A}{n_V^{1/3}}$ which does satisfy all assumptions. The standard deviation and interquartile range were calculated over subjects in the validation set on the difference in continuous covariate values (numerator of expression within $\phi$ in estimated likelihood and variance estimate).

We used total sample size of $n = 500$ and conducted 500 repetitions of the simulation for each set of parameter values. For each simulation, we used the proposed method to calculate estimates of the log hazard ratio, $\hat{\beta}$. We also calculated complete-case estimate using only true endpoints in the validation set, the naïve estimate using only uncertain endpoints from all subjects, and the true estimate using true endpoints from all subjects (which would be unavailable in real data). For each of the standard estimators, we used the maximum likelihood estimate rather than a partial likelihood estimate to better compare to our proposed method. We calculated estimated bias (parameter estimate$-$true parameter values), observed sample standard deviations (SD), estimated standard errors ($\widehat{\text{SE}}$), relative efficiency (RE) compared to the true estimator (where lower RE implies greater efficiency and RE equal to 1 implies optimal efficiency), mean squared error (MSE) estimates, and 95% coverage (Cov) at each of the observed time points. For all simulations presented in Tables 3.1 and 3.2, the observed sample standard deviation corresponds well with the standard error estimates from the asymptotic theory for the proposed estimator.

Table 3.1 shows the results from the simulations for a binary covariate. The log hazard ratio estimates estimated by our proposed method and the complete-case estimator are always unbiased, whereas the naïve estimates are biased whenever the correlation between outcomes is less than

1. Our proposed estimator has similar standard errors compared to the complete-case estimator when the correlation between outcomes is low. However, as the correlation between outcomes increases, our proposed estimator is able to incorporate more information from the non-validation set subjects and therefore improves in efficiency. This behavior is similar to what was seen in the nonparametric case.

Table 3.1: Simulation Results for Type 1 Censoring and a Binary Covariate

| $r$ | $\rho$ | Method | Bias $\times 10^{-3}$ | SD | $\widehat{\text{SE}}$ | MSE $\times 10^{-3}$ | RE | Cov |
|---|---|---|---|---|---|---|---|---|
| | | Proposed | 0.013 | 0.124 | 0.119 | 0.016 | 1.44 | 0.94 |
| | 0.01 | Comp | 0.010 | 0.123 | 0.119 | 0.015 | 1.42 | 0.95 |
| | | Naïve | -1.115 | 0.697 | 0.912 | 1.731 | 45.47 | 0.92 |
| | | Proposed | 0.013 | 0.124 | 0.119 | 0.016 | 1.44 | 0.94 |
| | 0.25 | Comp | 0.010 | 0.123 | 0.119 | 0.015 | 1.42 | 0.95 |
| | | Naïve | -0.510 | 0.251 | 0.246 | 0.323 | 5.89 | 0.41 |
| | | Proposed | 0.012 | 0.122 | 0.117 | 0.015 | 1.39 | 0.94 |
| 25 | 0.50 | Comp | 0.010 | 0.123 | 0.119 | 0.015 | 1.42 | 0.95 |
| | | Naïve | -0.445 | 0.170 | 0.163 | 0.227 | 2.70 | 0.23 |
| | | Proposed | 0.008 | 0.115 | 0.113 | 0.013 | 1.24 | 0.95 |
| | 0.75 | Comp | 0.010 | 0.123 | 0.119 | 0.015 | 1.42 | 0.95 |
| | | Naïve | -0.223 | 0.124 | 0.116 | 0.065 | 1.44 | 0.51 |
| | | Proposed | 0.007 | 0.106 | 0.103 | 0.011 | 1.04 | 0.94 |
| | 1.00 | Comp | 0.010 | 0.123 | 0.119 | 0.015 | 1.42 | 0.95 |
| | | Naïve | 0.004 | 0.103 | 0.103 | 0.011 | 1.00 | 0.94 |
| | | Proposed | 0.038 | 0.158 | 0.146 | 0.026 | 2.34 | 0.93 |
| | 0.01 | Comp | 0.026 | 0.153 | 0.147 | 0.024 | 2.18 | 0.95 |
| | | Naïve | -1.105 | 0.715 | 0.917 | 1.733 | 47.81 | 0.92 |
| | | Proposed | 0.040 | 0.160 | 0.146 | 0.027 | 2.40 | 0.92 |
| | 0.25 | Comp | 0.026 | 0.153 | 0.147 | 0.024 | 2.18 | 0.95 |
| | | Naïve | -0.510 | 0.251 | 0.246 | 0.323 | 5.89 | 0.41 |
| | | Proposed | 0.033 | 0.154 | 0.143 | 0.025 | 2.21 | 0.93 |
| 50 | 0.50 | Comp | 0.026 | 0.153 | 0.147 | 0.024 | 2.18 | 0.95 |
| | | Naïve | -0.445 | 0.170 | 0.163 | 0.227 | 2.70 | 0.23 |
| | | Proposed | 0.020 | 0.137 | 0.130 | 0.019 | 1.76 | 0.94 |
| | 0.75 | Comp | 0.026 | 0.153 | 0.147 | 0.024 | 2.18 | 0.95 |
| | | Naïve | -0.223 | 0.124 | 0.116 | 0.065 | 1.44 | 0.51 |
| | | Proposed | 0.016 | 0.110 | 0.103 | 0.012 | 1.13 | 0.93 |
| | 1.00 | Comp | 0.026 | 0.153 | 0.147 | 0.024 | 2.18 | 0.95 |
| | | Naïve | 0.004 | 0.103 | 0.103 | 0.011 | 1.00 | 0.94 |

$r$ is the percent missing and $\rho$ is the correlation between true and uncertain outcomes. Proposed refers to the proposed estimator, Comp refers to the complete-case estimator, and Naïve refers to the naïve estimator. SD is standard deviation of estimates across simulations, $\widehat{\text{SE}}$ is estimated standard error of the estimate, MSE is mean squared error, RE is relative efficiency, Cov is 95% coverage, all averaged across time.

We also simulated data assuming random censoring and changed the amount of censoring by sampling true censoring times $C$ from a uniform distribution. We considered a small amount of censoring (approximately 17%) using $C \sim$Unif$[3, 4]$, a moderate amount of censoring (approximately 36%) using $C \sim$Unif$[1, 4]$, and a large amount of censoring (approximately 84%) using $C \sim$Unif$[1, 2]$. Uncertain censoring times were simulated by $C^* = C + \gamma$ where $\gamma \sim$Unif$[0, 2]$. The results of these random censoring simulations for the binary covariate are shown in Table 3.2. Similar to the results with type 1 censoring, our proposed estimator and complete-case estimators have little bias compared to the naïve estimator and our proposed estimator is more efficient than the complete-case estimator regardless of the amount of censoring.

Table 3.2: Simulation Results for Random Censoring and a Binary Covariate

| $r$ | $C$ | Method | Bias $\times 10^{-3}$ | SD | $\hat{\text{SE}}$ | MSE $\times 10^{-3}$ | RE | Cov |
|---|---|---|---|---|---|---|---|---|
| | | Proposed | 0.007 | 0.114 | 0.112 | 0.013 | 1.17 | 0.94 |
| | 17 | Comp | -0.000 | 0.121 | 0.122 | 0.015 | 1.32 | 0.95 |
| | | Naïve | -0.145 | 0.110 | 0.107 | 0.033 | 1.09 | 0.73 |
| | | Proposed | 0.010 | 0.129 | 0.129 | 0.017 | 1.14 | 0.94 |
| 25 | 36 | Comp | 0.009 | 0.135 | 0.140 | 0.018 | 1.25 | 0.96 |
| | | Naïve | -0.125 | 0.132 | 0.123 | 0.033 | 1.19 | 0.80 |
| | | Proposed | 0.005 | 0.161 | 0.157 | 0.026 | 1.18 | 0.95 |
| | 84 | Comp | 0.004 | 0.165 | 0.166 | 0.027 | 1.25 | 0.95 |
| | | Naïve | -0.132 | 0.149 | 0.153 | 0.040 | 1.01 | 0.86 |
| | | Proposed | 0.023 | 0.129 | 0.124 | 0.017 | 1.50 | 0.93 |
| | 17 | Comp | 0.004 | 0.152 | 0.151 | 0.023 | 2.08 | 0.94 |
| | | Naïve | -0.145 | 0.110 | 0.107 | 0.033 | 1.09 | 0.73 |
| | | Proposed | 0.026 | 0.151 | 0.147 | 0.023 | 1.56 | 0.96 |
| 50 | 36 | Comp | 0.022 | 0.169 | 0.173 | 0.029 | 1.95 | 0.96 |
| | | Naïve | -0.125 | 0.132 | 0.123 | 0.033 | 1.19 | 0.80 |
| | | Proposed | 0.013 | 0.186 | 0.184 | 0.035 | 1.59 | 0.95 |
| | 84 | Comp | 0.003 | 0.201 | 0.205 | 0.040 | 1.85 | 0.95 |
| | | Naïve | -0.132 | 0.149 | 0.153 | 0.040 | 1.01 | 0.86 |

$r$ is the percent missing and $C$ is the percent censoring. Proposed refers to the proposed estimator, Comp refers to the complete-case estimator, and Naïve refers to the naïve estimator. SD is standard deviation of estimates across simulations, $\hat{\text{SE}}$ is estimated standard error of the estimate, MSE is mean squared error, RE is relative efficiency, Cov is 95% coverage, all averaged across time.

Results from simulations using the continuous covariate are forthcoming.

## 3.5. Data Example: Time to Development of Alzheimer's Disease

We illustrated our proposed semiparametric estimated likelihood method by considering data (retrieved on July 26, 2013) from the ongoing ADNI study (Weiner et al., 2012). Participants in this study were seen every 6 months until the end of two years, then annually thereafter, at which time clinical diagnoses of non-AD (cognitively normal or MCI) or AD were assessed. These follow-up times were predetermined by study design, and thus discrete survival estimates would be appropriate in this study. The current study includes data from participants in the ADNI-1 and ADNI-GO segments of the ADNI study. For those who agreed to a lumbar puncture, CSF assays were performed and $A\beta$ protein concentrations were measured. Participants with an $A\beta$ biomarker value greater than 192 pg/ml were classified as non-AD at baseline and those with an $A\beta$ value less than or equal to 192 pg/ml were classified as AD at baseline (Shaw et al., 2009). There were 186 patients who were non-AD at the time of enrollment according to both the clinical diagnosis and CSF diagnosis. For each patient, the time to clinical AD or last follow-up was recorded to obtain an uncertain, mismeasured outcome on all patients. A subset of 110 patients continued to have CSF assays performed annually. For these 110 patients in the validation set, patients were classified as non-AD or AD at each time point using the same cutoff of 192 pg/ml and the true time to AD or last follow-up was also recorded. Thus, $n_V = 110$ and $n = 186$. All patients also have information on gender and years of education.

We estimated the log hazard ratio, $\hat{\beta}$ of AD in females compared to males and comparing a one-year increase in education using our proposed method, the complete-case estimator using only 110 CSF diagnoses, and the naïve estimator using only 186 clinical diagnoses. For the complete-case and naïve estimators, we conducted estimation using both the maximum likelihood method and the more widely used partial likelihood method. For the partial likelihood method, we used Efron's approximation for ties (Efron, 1977). Table 3.3 shows the log hazard ratio and standard error estimates for both covariates. Both our proposed estimator and the complete-case estimator found a small positive log hazard ratio comparing females to males, which is similar to some literature indicating higher incidence of AD in women (Mielke et al., 2014; Andersen et al., 1999; Letenneur et al., 1999; Fratiglioni et al., 1997; Ott et al., 1998). However, the naïve estimate is large and negative. In this particular example comparing genders, the estimated standard errors from our proposed method and the complete-case method were similar. All estimators found a small positive

log hazard ratio of AD for a one-year increase in education, the direction of which initially appears to be inconsistent with the literature (Stern et al., 1994; Lindsay et al., 2002; Qiu et al., 2001; Sattler et al., 2012; Letenneur et al., 1999). However, previous studies often compared very low levels of education (i.e., less than eight years) to higher levels, whereas our sample had a mean of 16 years of education with very little variability. Therefore, along with the fact that the effect we observed was not significant, definitive conclusions about the true effect of education on time to AD cannot be made based on the current data. Nonetheless, our proposed estimator had a smaller estimated standard error than any other method, demonstrating improvements in efficiency.

Table 3.3: Data Example Log Hazard Ratio and Standard Error Estimates

| | Proposed | Comp (MLE) | Comp (Partial) | Naïve (MLE) | Naïve (Partial) |
|---|---|---|---|---|---|
| **Female** | | | | | |
| $\hat{\beta}$ | 0.306 | 0.164 | 0.218 | -1.609 | -1.689 |
| SE | 0.630 | 0.610 | 0.556 | 0.839 | 0.775 |
| **Education** | | | | | |
| $\hat{\beta}$ | 0.114 | 0.035 | 0.034 | 0.093 | 0.092 |
| SE | 0.070 | 0.118 | 0.103 | 0.089 | 0.102 |

Proposed refers to the proposed estimator, Comp refers to the complete-case estimator, and Naïve refers to the naïve estimator. MLE refers to estimated using the maximum likelihood method and Partial refers to estimated using the partial likelihood method. SE is the estimated standard error.

## 3.6. Discussion

We extended the nonparametric estimated likelihood method for data with uncertain endpoints and an internal validation subsample to the proportional hazards model with a binary or continuous covariate. The continuous covariate required a smooth kernel function to estimate probability distribution functions. Our proposed semiparametric method is consistent and asymptotically normal. Through simulation studies, we found that our proposed covariate effect parameter estimate is unbiased and its variance decreases as correlation between the uncertain and true outcome increases. By incorporating both uncertain and true endpoints in estimation, the proposed estimator can outperform both complete-case and naïve estimators.

As in the nonparametric case, we found that our proposed estimator behaves similarly to the complete-case estimator when the correlation between the uncertain and true outcomes is low. As correlation increases, the non-validation set subjects contribute more information and therefore

decrease variances of parameter estimates by providing more power. When correlation between outcomes is 1, or when the uncertain outcome has no measurement error, then our proposed estimator reduces to the maximum likelihood estimate based on complete true outcomes (no missingness).

In our current study, we evaluated the use of an estimated likelihood method with a single binary or a single continuous covariate. The method can easily be extended to consider multiple covariates, which would be useful in order to adjust for confounding variables or to consider categorical variables with more than two levels. Further study on the number of allowable covariates is warranted; however, based on the events per variable (EPV) testing in Chapter 2, it is expected that a similar EPV of 4 would apply to multivariate models. In these cases, the parameter vector would have dimension equal to the number of coefficients (or covariates) plus the number of time points for the event distribution plus the number of time points for the censoring distribution if there is random censoring.

Interesting study design issues arise with regard to the size of the validation sample that is needed to adequately accommodate the uncertainty of the mismeasured endpoints. For example, in clinical trials that aim to evaluate a treatment effect, it is valuable to determine the optimum size of the study cohort and optimum size of the validation subsample to achieve a pre-specified power. These design issues are discussed in Chapter 4.

Due to the difficulty in obtaining true outcomes on many subjects, the methods we have proposed have useful applications in clinical trials. Designing studies such that uncertain outcomes are collected on all patients and true outcomes only collected on a subsample of patients can save on trial costs and ensure that an adequate number of patients are enrolled. Using our proposed semiparametric estimated likelihood method to analyze these data can provide accurate and powerful statistical inference to evaluate treatment effects.

# CHAPTER 4

# OPTIMAL STUDY DESIGN FOR ASSESSING TREATMENT EFFECTS IN TIME-TO-EVENT DATA WITH UNCERTAIN ENDPOINTS AND A VALIDATION SUBSAMPLE

## 4.1. Introduction

In clinical trials involving time-to-event data, many outcomes of interest are too cost-prohibitive or invasive to obtain on a large sample of patients. Often, alternative outcomes that measure the true outcome with some error can be used to increase the total sample size. In these situations where mismeasured, uncertain outcomes may be available on many subjects and true outcomes are only available on a smaller subsample, there are new methods that have been developed to estimate the hazard ratio of a treatment effect (see Chapter 3). It is important to develop optimal design strategies under circumstances with uncertain endpoints and an internal validation subsample when using these methods of survival analysis.

There are several methods that have been developed for computing optimal sample sizes required for survival analysis studies (Freedman, 1982; Lakatos, 1986, 1988; Schoenfeld, 1981, 1983; Shih, 1995). Freedman (1982) and Schoenfeld (1981; 1983) proposed simple sample size formulas that are in wide use today, but their formulas do not take into account potential mismeasurement of uncertain outcomes and cannot incorporate both an uncertain outcome on all subjects and true outcomes on a subsample of subjects.

The methods developed in Chapter 3 rely on estimated likelihood functions, where the conditional probability of the uncertain outcome given the true outcome is estimated. As the correlation between the uncertain and true outcome increases, the uncertain outcome is able to provide more information and therefore power to the parameter estimate of interest. Furthermore, it is not only the total sample size that drives the power and variance estimates of the survival function estimator, but also the size of the validation set, where the validation set is the subsample of patients for whom both the uncertain and true outcomes are available.

In this paper, we develop optimal design strategies for a range of study conditions, including varying

effect sizes, correlations between outcomes, percentage of missing true outcomes, and baseline distributions. We use simulations to calculate sample sizes given a pre-specified power of a Wald-type test for detecting a difference across treatment groups and describe the steps in Section 4.2. We also propose a sample size formula in Section 4.3 based on pre-specified power, effect size, correlation between outcomes, and percent missingness. Finally, we compare the advantages and disadvantages of each method in Section 4.5 and discuss future directions.

## 4.2. Sample Size Calculation through Simulations

We conducted simulations to calculate the total number of true events in the total sample and in the validation set needed to detect a log hazard ratio between two equally-sized treatment groups when using the estimated likelihood method developed in Chapter 3. The parameters needed to conduct the simulations include a baseline distribution of the true event, $F_0$, the log hazard ratio of interest, $\beta$, the correlation between the uncertain and true outcomes, $\rho$, the proportion of true outcomes that are missing, $r$, and the desired power. We assume $M$ simulation repetitions will be conducted.

First, the probability distribution function of the true event in each treatment group is calculated by using the baseline distribution $F_0$ and effect size assuming proportional hazards between groups over time. We assume positive, uniformly distributed measurement error and use the correlation between outcomes to determine the conditional distribution of uncertain outcomes given true outcomes and treatment group as in Chapter 3. For a given even-valued total number of events, $d$, the proportion of missingness is used to calculate the number of events in the validation set by first calculating $d \cdot (1 - r)$, then rounding this value to the smallest even integer greater than $d \cdot (1 - r)$, $d_V$. This ensures that the treatment groups are also equal in size among the validation set as well as the total sample. Without loss of generality, a binary validation set indicator is used to mark the first $d - d_V$ subjects as subjects in the non-validation set and the rest of the $d_V$ subjects as those in the validation set.

For each simulation repetition, a treatment indicator, $Z$, is evenly distributed across the validation set and total sample. Then based on the value of the treatment indicator, event times are randomly sampled from the probability distribution function of the true event given treatment group. However, if the provided baseline survival function reaches 0 at the last time point, then there will be a

boundary value problem with the Newton-Raphson maximization algorithm. In this case, $d$ event times are randomly sampled from all time points except the last time point. Then, the data are augmented by events at the last time point and these extra observations are considered censored at the second to last time point. All other event indicator values will be 1, such that there are always $d$ total true events. Given each subject's true event times and treatment indicator value, uncertain event times are sampled from the conditional distribution of uncertain event times given true outcome and treatment group. All uncertain event indicators before the last time point are given value 1 and those with value at the last time point are considered censored at the second to last time point. Finally, the true event times and indicators for subjects in the non-validation set are removed to be missing.

Now that a full dataset has been generated, the maximization algorithm and variance calculation from Chapter 3, Section 3.2 can be utilized to estimate $\hat{\beta}$ and $\mathrm{Var}(\hat{\beta})$. We conduct a two-sided Wald-type test of the null hypothesis $H_0 : \beta = 0$ and record whether the test is rejected or not rejected. This procedure is repeated for all simulation repetitions and the proportion of rejections out of $M$ repetitions is calculated. This proportion represents the calculated power of the test given $d$ total true events and $d_V$ true events in the validation set.

To determine optimal sample size, a binary search algorithm is used. The algorithm is initiated at the value given by Schoenfeld's (1983) sample size formula for a standard Cox proportional hazards model, $d_S$. Following the procedure described above, the calculated power from $M$ simulation repetitions is compared to the pre-specified power, for example, 0.80. One of two situations can occur:

1. If the calculated power is greater than or equal to 0.80, then the next sample size to be tested is halfway between the minimum sample size and the current sample size. Since the minimum sample size after the first iteration of the binary search algorithm is 2, then the next sample size to be tested is $(d_S + 2)/2$ (or $(d_S + 2)/2 + 1$ if $(d_S + 2)/2$ is odd). The new maximum sample size would then become $d_S$.

2. If the calculated power is less than 0.80, then the next sample size to be tested is halfway between the maximum sample size and the current sample size. Since there is no maximum sample size after the first iteration of the binary search algorithm, we make the maximum

equal to 40 more than the current size. Then the next sample size to be tested is $(d_S + d_S + 40)/2$ (or $(d_S + d_S + 40)/2 + 1$ if $(d_S + d_S + 40)/2$ is odd). The new minimum sample size would then become $d_S$.

These steps are repeated until the previous and current sample sizes being tested are within 2 of each other. The optimal sample size is then considered to be the smallest sample size such that the calculated power is within the interval (0.795,0.820).

Using simulations, we calculated optimal sample sizes for baseline distribution $T \sim$Unif$[1, 5]$, correlations of $\rho \in \{0.25, 0.50, 0.75, 1.00\}$, proportions of missingness of $r \in \{0, 0.25, 0.50\}$, effect sizes of $\beta \in \{0.41, 0.50, 0.69\}$, and desired power of 0.80. An example is given in Figure 4.1 for $\beta = 0.50$ to show the total number of true events in the sample ($d$, solid lines) and number of true events in the validation set ($d_V$, dashed lines). We found that the optimal total number of true events and number of true events in the validation set was similar to the optimal number of events calculated by Schoenfeld's (1983) sample size formula when there was no missingness. When missingness was greater than 0 and correlation between outcomes was low, the number of true events in the validation set was similar to the standard value. As correlation between outcomes increased, the number of true events in the validation set decreased. At perfect correlation of $\rho = 1$, the total number of true events in the sample was similar to the standard value.

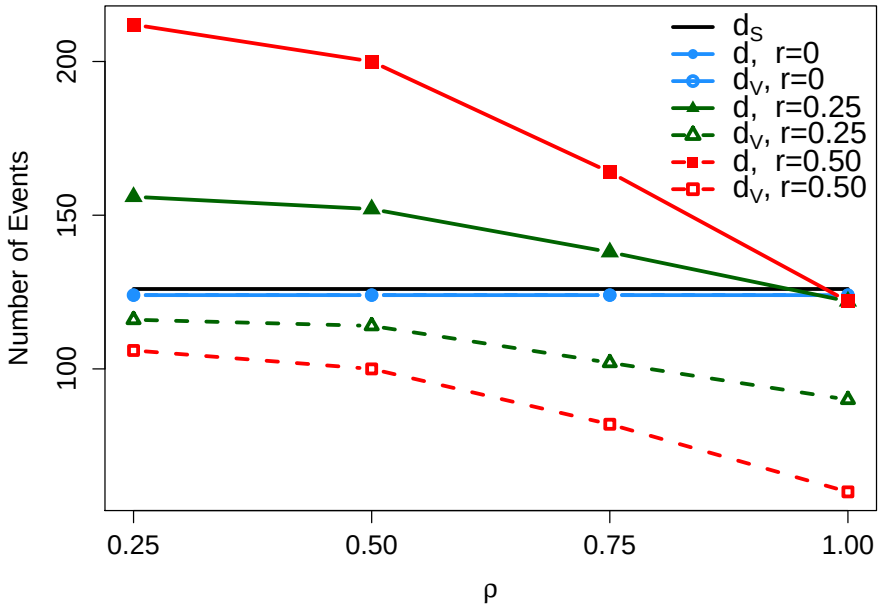
Although our simulation results give both total number of true events in the sample and number of true events in the validation set, the number of true events in the non-validation set would not be observed in a real study. Therefore, it would be most relevant to use the number of true events in the validation set for planning a trial. In order to calculate the total number of subjects to recruit, we can use the missingness proportion $r$ as well as an assumed constant proportion of censoring, $c$. The total number of subjects needed to be recruited in the study would be

$$n = \left\lceil \left\lceil \frac{d_V}{(1-r)(1-c)} \right\rceil \right\rceil \tag{4.1}$$

and the number of subjects needed for the validation set would be

$$n_V = \left\lceil \left\lceil \frac{d_V}{1-c} \right\rceil \right\rceil \tag{4.2}$$

Figure 4.1: Optimal Number of True Events for $T \sim$Unif$[1, 5]$ and $\beta = 0.50$




This figure appears in color in the electronic version of this article.

where $\lceil\lceil a \rceil\rceil$ here represents the smallest *even* integer greater than $a$. The even value ensures that the two treatment groups can have equal sizes in the total sample and in the validation set. The value of $c$ can either be assumed or estimated according to the study design, such as by using the steps in Section 2.1 of Schoenfeld (1983) for estimating the proportion of patients that will have an event.

Although the estimated likelihood method sometimes requires a greater number of total subjects compared to the standard method, it requires a similar or smaller number of true outcomes. Therefore, when the true outcomes are expensive or otherwise difficult to obtain, the estimated likelihood method can save on costs without sacrificing power. This is particularly true when there is high correlation between the uncertain and true outcomes.

Computing optimal sample sizes by conducting simulations allows great flexibility in the specification of study parameters. However, running the simulations is time-consuming, especially when $M$ is large and when standard errors must be calculated by bootstrap samples. An ideal alternative would be to use a sample size formula.

## 4.3. Sample Size Formula

Consider testing the null hypothesis $H_0 : \beta = \beta_0$, against the alternative of $H_a : \beta = \beta_a$. Then our effect size of interest is $\beta_a - \beta_0$. We adapted Schoenfeld's (1983) sample size formula by multiplying by relative efficiency of the estimated log hazard ratio based on our proposed method to the true estimate using complete data with no missing true outcomes. The number of true events in the validation set can be calculated by

$$d_V = \left\lceil \left\lceil 4 \cdot \frac{(z_{1-\alpha/2} - z_{1-\text{Power}})^2}{(\beta_a - \beta_0)^2} \left( \frac{1}{1-r} - \frac{r}{1-r}\rho^2 \right) (1-r) \right\rceil \right\rceil . \tag{4.3}$$

where $z_\alpha = \Phi^{-1}(\alpha)$ is the $\alpha$ percentile of the normal distribution. We can then use Equations 4.1 and 4.2 to calculate the total number of subjects and number in the validation set needed to be recruited for the trial.

The relative efficiency expression is a function of proportion missingness $r$ and correlation $\rho$ between the true and uncertain outcomes. The concave shape is based on observations from simulations in previous work (Chapter 3). Specifically, our proposed estimator reduces to the complete-case estimator for a useless uncertain outcome, so the relative efficiency of the proposed estimator at $\rho = 0$ is equal to the relative efficiency of the complete-case estimator for a given percent of missingness, $RE_{max}$. Similarly, for the case of a perfect uncertain outcome with no measurement error, the relative efficiency of the proposed estimator at $\rho = 1$ is 1. From previously conducted simulations, the relative efficiency begins decreasing slowly as correlation increases, then falls quickly as correlation approaches 1. Therefore, we write the relative efficiency as

$$RE = RE_{max} - (RE_{max} - 1) \cdot \rho^2. \tag{4.4}$$

Although Schoenfeld's sample size formula is based on a score test, it can be re-written as a Wald test. When written as a Wald test, the variance component of the estimator is $4/d$, where $d$ is the number of events. For the complete-case estimator, the number of events is the number of events in the validation subsample, or $(1-r)n$. For the true estimator, the number of events is the total

sample size, $n$. Therefore, the maximum relative efficiency is

$$RE_{max} = \frac{4/\{(1-r)n\}}{4/n} = \frac{1}{1-r} \tag{4.5}$$

and the relative efficiency equation can be written as

$$RE = \frac{1}{1-r} - \frac{r}{1-r}\rho^2. \tag{4.6}$$

This equation multiplied by Schoenfeld's sample size formula and $(1-r)$ for the proportion in the validation set gives Formula 4.3.

We tested the performance of our proposed sample size formula by calculating power based on the proposed formula and determining power by Monte Carlo simulations for several sample sizes. Simulations are based on data sampled from baseline distribution $T \sim$Unif$[1, 5]$, correlations of $\rho \in \{0.25, 0.50, 0.75, 1.00\}$, proportions of missingness of $r \in \{0, 0.25, 0.50\}$, and effect sizes of $\beta \in \{0.41, 0.50, 0.69\}$. Results are shown in Table 4.1 for values of $d_V$ up to 100 and in Table 4.2 for values of $d_V$ up to 200. The power calculations are mostly similar, indicating that the proposed sample size formula can give good estimates of the optimal sample size. In some cases, the formula slightly underestimates power, but this is the preferred direction in that it would slightly overestimate optimal sample sizes needed.

Although the sample size formula performs well compared to these simulations, further work to test its robustness under different study parameters is under way. We will use a baseline distribution of $T \sim$Unif$[1, 8]$ to demonstrate a different number of time points, baseline distribution of $T \sim$Geometric$(0.5)$ to demonstrate a different discrete distribution, and $\epsilon \sim$Geometric$(0.5)$ to demonstrate a different measurement error distribution, results of which are forthcoming.

## 4.4. Example

The Ginkgo Evaluation of Memory (GEM) study was a randomized controlled trial with the primary objective of comparing *Ginkgo biloba* extract to a placebo for prevention of all-cause dementia (DeKosky et al., 2006). A retrospective secondary study using the GEM data evaluated several antihypertensive drugs and found that diuretics, angiotensin-1 receptor blockers (ARB), and angiotensin-converting enzyme inhibitors (ACE-I) were associated with reduced risk of Alzheimer's

Table 4.1: Power Estimated by Formula 4.3 and by Simulations for $d_V$ up to 100

| $d_V$ | $\beta$ | $r$ | $\rho$ | Power (Formula) | Power (Monte Carlo) |
|---|---|---|---|---|---|
| 100 | 0.41 | 0.00 | 0.25 | 0.536 | 0.530 |
| 100 | 0.41 | 0.00 | 0.50 | 0.536 | 0.530 |
| 100 | 0.41 | 0.00 | 0.75 | 0.536 | 0.530 |
| 100 | 0.41 | 0.00 | 1.00 | 0.536 | 0.530 |
| 76 | 0.41 | 0.25 | 0.25 | 0.437 | 0.450 |
| 76 | 0.41 | 0.25 | 0.50 | 0.454 | 0.458 |
| 76 | 0.41 | 0.25 | 0.75 | 0.485 | 0.478 |
| 76 | 0.41 | 0.25 | 1.00 | 0.536 | 0.534 |
| 50 | 0.41 | 0.50 | 0.25 | 0.313 | 0.336 |
| 50 | 0.41 | 0.50 | 0.50 | 0.341 | 0.368 |
| 50 | 0.41 | 0.50 | 0.75 | 0.401 | 0.418 |
| 50 | 0.41 | 0.50 | 1.00 | 0.536 | 0.560 |
| 100 | 0.50 | 0.00 | 0.25 | 0.705 | 0.676 |
| 100 | 0.50 | 0.00 | 0.50 | 0.705 | 0.676 |
| 100 | 0.50 | 0.00 | 0.75 | 0.705 | 0.676 |
| 100 | 0.50 | 0.00 | 1.00 | 0.705 | 0.676 |
| 76 | 0.50 | 0.25 | 0.25 | 0.593 | 0.566 |
| 76 | 0.50 | 0.25 | 0.50 | 0.613 | 0.586 |
| 76 | 0.50 | 0.25 | 0.75 | 0.649 | 0.620 |
| 76 | 0.50 | 0.25 | 1.00 | 0.705 | 0.680 |
| 50 | 0.50 | 0.50 | 0.25 | 0.435 | 0.442 |
| 50 | 0.50 | 0.50 | 0.50 | 0.472 | 0.458 |
| 50 | 0.50 | 0.50 | 0.75 | 0.550 | 0.520 |
| 50 | 0.50 | 0.50 | 1.00 | 0.705 | 0.698 |
| 100 | 0.69 | 0.00 | 0.25 | 0.932 | 0.914 |
| 100 | 0.69 | 0.00 | 0.50 | 0.932 | 0.914 |
| 100 | 0.69 | 0.00 | 0.75 | 0.932 | 0.914 |
| 100 | 0.69 | 0.00 | 1.00 | 0.932 | 0.914 |
| 76 | 0.69 | 0.25 | 0.25 | 0.858 | 0.836 |
| 76 | 0.69 | 0.25 | 0.50 | 0.873 | 0.842 |
| 76 | 0.69 | 0.25 | 0.75 | 0.899 | 0.894 |
| 76 | 0.69 | 0.25 | 1.00 | 0.932 | 0.916 |
| 50 | 0.69 | 0.50 | 0.25 | 0.698 | 0.692 |
| 50 | 0.69 | 0.50 | 0.50 | 0.742 | 0.704 |
| 50 | 0.69 | 0.50 | 0.75 | 0.821 | 0.802 |
| 50 | 0.69 | 0.50 | 1.00 | 0.932 | 0.918 |

Power (Formula) refers to the power calculated by the proposed sample size formula and Power (Monte Carlo) refers to the power calculated through simulations.

Table 4.2: Power Estimated by Formula 4.3 and by Simulations for $d_V$ up to 200

| $d_V$ | $\beta$ | $r$ | $\rho$ | Power (Formula) | Power (Monte Carlo) |
|---|---|---|---|---|---|
| 200 | 0.41 | 0.00 | 0.25 | 0.826 | 0.834 |
| 200 | 0.41 | 0.00 | 0.50 | 0.826 | 0.834 |
| 200 | 0.41 | 0.00 | 0.75 | 0.826 | 0.834 |
| 200 | 0.41 | 0.00 | 1.00 | 0.826 | 0.834 |
| 150 | 0.41 | 0.25 | 0.25 | 0.716 | 0.734 |
| 150 | 0.41 | 0.25 | 0.50 | 0.737 | 0.732 |
| 150 | 0.41 | 0.25 | 0.75 | 0.773 | 0.790 |
| 150 | 0.41 | 0.25 | 1.00 | 0.826 | 0.834 |
| 100 | 0.41 | 0.50 | 0.25 | 0.549 | 0.538 |
| 100 | 0.41 | 0.50 | 0.50 | 0.592 | 0.550 |
| 100 | 0.41 | 0.50 | 0.75 | 0.677 | 0.626 |
| 100 | 0.41 | 0.50 | 1.00 | 0.826 | 0.834 |
| 200 | 0.50 | 0.00 | 0.25 | 0.942 | 0.928 |
| 200 | 0.50 | 0.00 | 0.50 | 0.942 | 0.928 |
| 200 | 0.50 | 0.00 | 0.75 | 0.942 | 0.928 |
| 200 | 0.50 | 0.00 | 1.00 | 0.942 | 0.928 |
| 150 | 0.50 | 0.25 | 0.25 | 0.870 | 0.872 |
| 150 | 0.50 | 0.25 | 0.50 | 0.885 | 0.874 |
| 150 | 0.50 | 0.25 | 0.75 | 0.910 | 0.882 |
| 150 | 0.50 | 0.25 | 1.00 | 0.942 | 0.928 |
| 100 | 0.50 | 0.50 | 0.25 | 0.719 | 0.738 |
| 100 | 0.50 | 0.50 | 0.50 | 0.762 | 0.744 |
| 100 | 0.50 | 0.50 | 0.75 | 0.839 | 0.784 |
| 100 | 0.50 | 0.50 | 1.00 | 0.942 | 0.928 |
| 200 | 0.69 | 0.00 | 0.25 | 0.998 | 1.000 |
| 200 | 0.69 | 0.00 | 0.50 | 0.998 | 1.000 |
| 200 | 0.69 | 0.00 | 0.75 | 0.998 | 1.000 |
| 200 | 0.69 | 0.00 | 1.00 | 0.998 | 1.000 |
| 150 | 0.69 | 0.25 | 0.25 | 0.989 | 0.986 |
| 150 | 0.69 | 0.25 | 0.50 | 0.992 | 0.982 |
| 150 | 0.69 | 0.25 | 0.75 | 0.995 | 0.992 |
| 150 | 0.69 | 0.25 | 1.00 | 0.998 | 1.000 |
| 100 | 0.69 | 0.50 | 0.25 | 0.939 | 0.944 |
| 100 | 0.69 | 0.50 | 0.50 | 0.958 | 0.944 |
| 100 | 0.69 | 0.50 | 0.75 | 0.983 | 0.966 |
| 100 | 0.69 | 0.50 | 1.00 | 0.998 | 1.000 |

Power (Formula) refers to the power calculated by the proposed sample size formula and Power (Monte Carlo) refers to the power calculated through simulations.

disease (AD) among patients who had normal cognition at baseline, but only diuretics were significantly associated with reduced risk of AD among those with mild cognitive impairment (MCI) (Yasar et al., 2013). Yasar et al. (2013) suggested that the lack of significant associations with other antihypertensives was due to lack of power, since hazard ratios showed trends for an effect and only 110 out of 320 patients who had MCI at baseline eventually developed AD. Specifically, hazard ratios comparing antihypertensive medication use to no antihypertensive medication use from adjusted Cox proportional hazards models were 0.38 for diuretics, 0.37 for ARB, and 0.53 for ACE-I (Yasar et al., 2013).

Suppose that investigators were interested in conducting new, separate randomized controlled trials to compare diuretics, ARB, and ACE-I against no antihypertensive use among patients with MCI to evaluate effects on development of AD. In order to obtain accurate and efficient hazard ratio estimates, the methods developed by Zee and Xie (Chapter 3) can be used after obtaining an uncertain, clinically diagnosed time to AD outcome on all patients and a true time to pathological AD outcome on a validation subset of patients, measured based on cerebral spinal fluid (CSF) assays of $A\beta$ protein concentrations (Shaw et al., 2009). Suppose that the non-validated set was predicted to be 41% of the total sample, as in the Alzheimer's Disease Neuroimaging Initiative (ADNI) data from Section 3.5. Using the validation set from the same ADNI data, we estimated the correlation between true and uncertain observed times to be 0.788.

We assumed values $r = 0.41$, $\rho = 0.788$, power of 0.80, and log hazard ratios of $\beta_a = \log(0.38) = -0.968$ for diuretics, $\beta_a = \log(0.37) = -0.994$ for ARB, and $\beta_a = \log(0.53) = -0.635$ for ACE-I. For two-sided Wald tests at significance level $\alpha = 0.05$ and null hypothesis of $H_0 : \beta_0 = 0$, the numbers of true events in the validation set, $d_V$, calculated by Formula 4.3 are shown in Table 4.3, along with the number of true events needed if using standard methods, $d_S$. Using a censoring proportion of $c = 210/320 = 0.656$ based on the proportion observed in the GEM study, we also calculated the total number of subjects, $n$, and number of subjects in the validation set, $n_V$, that would need to be recruited using Equations 4.1 and 4.2. Finally, we used Equation 4.2 but substituted $d_S$ for $d_V$ to calculate $n_S$, the number of subjects that would need to be recruited if using standard survival methods.

The example results show that the numbers of true events that need to be observed would be smaller if using the estimated likelihood method compared to the standard method ($d_V$ vs. $d_S$). The

Table 4.3: Optimal Number of Events in Study Design Example

| | Proposed Method | | | Standard | |
|---|---|---|---|---|---|
| Antihypertensive | $d_V$ | $n$ | $n_V$ | $d_S$ | $n_S$ |
| diuretic | 26 | 130 | 76 | 34 | 100 |
| ARB | 24 | 120 | 70 | 32 | 94 |
| ACE-I | 60 | 296 | 176 | 78 | 228 |

total number of subjects that would need to be recruited for the estimated likelihood method ($n$) is higher than the total number of subjects to be recruited for the standard method ($n_S$). However, there is a smaller number of subjects for which the true outcome must be collected ($n_V$ vs. $n_S$). For expensive or otherwise difficult to obtain true outcomes, this shows potential savings in cost or an increase in power when using the estimated likelihood method.

## 4.5. Discussion

We developed two methods for calculating the optimal number of true events in the validation set when using the estimated likelihood method developed by Zee and Xie (Chapter 3). We also demonstrated how to calculate the total number of subjects and number in the validation set that would need to be recruited for a trial. We found that the number of true outcomes needed is similar to that according to Schoenfeld's (1983) formula for using the standard proportional hazards model when there are no true outcomes missing or when the correlation between uncertain and true outcomes is low. As correlation between outcomes increases, the number of true outcomes needed decreases. Therefore, using the estimated likelihood method can save on costs by obtaining fewer true outcomes.

By simulating data and using a binary search algorithm, we calculated optimal sample sizes such that all study parameters can be controlled. However, this complex procedure takes a long time to complete. As an alternative, we have proposed a sample size formula that is easy and fast to use and gives similar results to those from simulations. Robustness checks and methods on further improving the precision of the sample size formula are currently under investigation.

The parameters required for the sample size formula include the effect size of interest measured as the log hazard ratio, $\beta$, the correlation between the uncertain and true outcome, $\rho$, the proportion of true outcomes that are missing, $r$, and the desired power. Power can be set at 0.80 or 0.90. For all other parameters, values can be obtained from pilot data or from the literature, as we did in

our data example. The simulation method requires that an additional parameter be specified: the baseline survival distribution of the true outcome, $F_0$. This parameter is one that would not typically be available from the literature. Therefore, it must be assumed or estimated from pilot data. This may be considered another drawback of using the simulation method.

Following either method for calculating optimal sample sizes should allow an investigator to design a new trial where uncertain outcomes will be collected on all participants and true outcomes collected on a subsample. We recommend using this data structure and the methods of Zee and Xie (Chapter 3) to analyze these data in order to obtain accurate and efficient survival estimates while saving on costs of a trial.

# CHAPTER 5

## CONCLUSION

In this dissertation, we developed methods to conduct survival analysis using data with uncertain outcomes on all subjects and true outcomes on a validation subsample. We also developed optimal study design strategies for new trials with these data characteristics. In Chapter 2, we developed a nonparametric discrete survival function estimator using an estimated likelihood method, derived from Pepe's (1992) framework for general outcomes. The likelihood allowed for fixed or random censoring mechanisms and allowed for any subject to be validated. Due to the discrete nature of the time points, conditional probabilities within the likelihood were estimated empirically using proportions based on observed data in the validation set. We maximized the likelihood function using a Nelder-Mead algorithm to calculate the maximum estimated likelihood estimator. We showed that the estimator is consistent and asymptotically normal, and we conducted a series of simulations to test the performance of our method. We found that the proposed estimator was unbiased, whereas the naïve Kaplan-Meier survival function estimator, which uses only uncertain outcomes on all subjects, was often biased. Our proposed estimator also behaved similarly to the complete-case Kaplan-Meier survival function estimator, which uses only true outcomes from the validation set, when the correlation between the true and uncertain outcomes was low. As the correlation increased, our proposed estimator was able to use more information from the non-validation set subjects to improve in efficiency. At perfect correlation between outcomes, or when the uncertain outcome had no measurement error, our proposed estimator reached optimal efficiency. We found these results whether we used fixed or random censoring mechanisms. Through our simulations, we found that these properties held when there was 50% or less missingness of the true outcomes and when there was an EPV of at least 4. Finally, we illustrated our method by calculating survival function estimates of the time to AD and standard errors using data from the ADNI study. In the data example, we saw that the survival function estimate calculated by our proposed method was similar to that of the complete-case Kaplan-Meier estimate, whereas the naïve Kaplan-Meier estimate was slightly separated from the other two curves. We also observed slight gains in efficiency when using our proposed method as compared to the complete-case estimate.

In Chapter 3, we extended our methods to the semiparametric case in order to assess the effects of

a binary or continuous covariate. We assumed a proportional hazards model and aimed to estimate the log hazard ratio of the event of interest for different values of the covariate. For a single binary covariate, the estimator and its variance was similar to the nonparametric case, but with the addition of a random variable and the proportional hazards model. For the continuous covariate, however, the empirically estimated probabilities based on indicator functions would lead to zero-valued contributions by many if not all non-validation set subjects. Therefore, we used a smooth kernel function and bandwidth in the likelihood and variance estimate when we had a continuous covariate. Two additional assumptions about the choice of bandwidth were required for the asymptotic properties to hold. With the assumptions in place, we showed that the estimate of the log hazard ratio for both the binary and continuous case was consistent and asymptotically normal. Using simulations, we compared our proposed semiparametric estimated likelihood method to the complete-case and naïve estimators based on the maximum likelihood method for the Cox proportional hazards model. We found similar results as in the nonparametric version: our proposed estimator was unbiased and was as or more efficient than the complete-case estimator. We illustrated the semiparametric methods by using the ADNI data to estimate the effect of gender (binary) and education (continuous) on time to AD. We observed that our proposed estimator of the log hazard ratio had similar standard error as the complete-case estimator for gender but lower standard error for education.

In Chapter 4, we developed optimal study design strategies to calculate the number of true events in the validation set that would need to be observed to achieve a pre-specified power when comparing survival across two groups. We did so using simulations, which allows great flexibility in specification of parameters but takes a long time to complete. As expected from the properties observed from our semiparametric estimated likelihood estimator, the optimal number of true events in the validation set was similar to the standard number of events when correlation between outcomes was low. As correlation increased, the optimal number of true events in the validation set decreased. We also proposed a sample size formula adapted from Schoenfeld's (1983) formula for the standard proportional hazards model, which is much easier to use than the simulations. By comparing power calculated by both methods, we found the sample size formula gives close but not exactly the same values as the simulations do. The parameters that need to be specified for both simulation and formula methods include effect size, correlation between outcomes, proportion of missingness, and desired power. Using simulations also requires that the baseline distribution of true event times be specified, which may be difficult to obtain. We demonstrated the use of our

proposed sample size formula by using parameters from a follow-up study to the GEM trial and from the ADNI data to estimate the number of true events that would need to be observed in new trials comparing use of three antihypertensive medications to no antihypertensive use on development of AD. We found that our proposed method would require fewer true events to be observed than the standard method would. We also calculated the number of total subjects and number of subjects in the validation set that would need to be recruited based on proportion missingness and a fixed proportion of censoring. Our example showed that our proposed method would require a larger total number of subjects to be recruited, but a smaller number of subjects on which to obtain the true outcome, as compared to the standard method.

Because true survival outcomes may be difficult to obtain for a large number of subjects, we recommend collecting an uncertain outcome on all subjects as well as an internal validation subsample of true outcomes. The uncertain outcomes can be surrogate or auxiliary markers that measure the true outcomes with some error and their mismeasurement rates do not have to be known or estimated. The correlation between uncertain and true outcomes is also unnecessary in order to use our proposed methods. We showed that the nonparametric and semiparametric estimated likelihood methods can be used to analyze discrete survival data with these characteristics to improve accuracy and power as compared to the standard methods. As more efficient (both in cost and in parameter estimation) clinical trials becomes increasingly necessary, we recommend using our proposed methods to save on trial costs without sacrificing power or to increase power without sacrificing costs.

## 5.1. Future Directions

### *5.1.1. Multivariable Models*

There are several interesting areas of future study to consider. The semiparametric methods that we developed are specifically designed for the situation where only a single binary or single continuous covariate is considered. It is natural to extend this to a multivariable model, in order to incorporate discrete, multi-level categorical covariates and to adjust estimates for confounding variables. Although the extension to multivariable modeling should be somewhat straightforward, there are a few additional details that must be examined.

First, the number of covariates that can be included in the model may need to be limited. As in standard regression modeling, there are limits on the number of parameters to allow in a model based on the sample size in order to avoid overfitting (Harrell Jr, 2001). In a semiparametric estimated likelihood method with multiple covariates, it is not only the number of covariates that have associated parameters but also the event and censoring survival functions that must be estimated. Based on our simulations studies in Chapter 2, we found that the events per variable (EPV) required for the nonparametric method was 4. It is expected that we would find a similar EPV when considering multivariable models, but this must be verified, particularly when considering continuous covariates.

Second, the form of the likelihood and associated asymptotic theory may need to be altered when considering complex models. For example, if potential effect modification must be evaluated, an interaction term may be included in the model. Specifically, consider an interaction term between years of education and female gender, which will have 0 values for all males. When calculating the likelihood contribution for a male non-validation set subject with 12 years of education, the zero-valued interaction term for this subject will seem close to an interaction term with value 1, which would correspond to a female with 1 year of education. The kernel function would then give a large value, implying that these two subjects are closely matched. The effects of such anomalies when considering interaction terms or other more complicated covariates in the estimated likelihood must be studied.

### *5.1.2. Sample Size Formula Improvements*

The sample size formula proposed in Section 4.3 produced power values close to those from simulations overall. However, some values were slightly lower. Therefore, additional work on improving the precision of the formula would be useful. To do so, we can explore the shape of the convex function relating relative efficiency to the correlation between outcomes. In the development of our proposed formula, we attempted to fit several different convex functions. We used both quartic and negative log functions before concluding that the quadratic function produced the closest relative efficiency values to the ones observed and did not suffer from boundary condition issues. Further research on the best fitting function may give an even more accurate formula.

It may also be beneficial to develop an alternative sample size formula that relies on fewer parameter specifications. For example, we currently assume that the proportion of missingness is fixed.

However, it may be possible to control the proportion of missingness in some clinical trials through recruitment efforts. In these cases, study design strategies that allow for a flexible proportion of missingness may be needed to calculate optimal sample sizes across varying levels of missingness. Rather than a one-dimensional problem of solving for only the number of true events in the validation set, we would also need to solve for the number of total events in the study simultaneously.

### *5.1.3. Continuous Survival Times*

The survival analysis and study design methods proposed in this dissertation rely on the assumption that study visit times are pre-determined. This results in a discrete survival time. However, some studies may have survival times measured on what would be considered a continuous scale. For example, if the time unit of interest is in days but patients cannot be evaluated every day, we can consider time to be continuous and the possible observed visit times are no longer pre-determined. The proposed estimated likelihood methods would not be useful in this case because the non-validation set subjects would often if not always contribute zero values to the likelihood. As we did for the continuous covariate, we could use a smooth kernel function instead of an indicator function in the empirical probability estimates for the continuous outcome. However, the infinite dimensional parameter space produced by the continuous time data poses a problem. Although the times can be discretized based on observed times, as can be done for the Kaplan-Meier estimator, the dimensionality of the parameter space is typically still large. Maximization of the estimated likelihood over a large number of parameters is not only slow to converge but may also fail to converge at all. Furthermore, assuming an EPV of about 4, the large number of parameters would require an even larger sample. In the case of continuous times, however, a larger sample typically implies a larger number of parameters. This endless cycle makes it difficult if not impossible to acquire a large enough sample size for parameter estimation with the estimated likelihood method.

Rather than using an estimated likelihood method to analyze continuous survival times when there is data on uncertain outcomes for all subjects and true outcomes for a subsample, another potential method is the mean score method developed by Pepe et al. (1994). The mean score method also has a validation set portion and a non-validation set portion. The validation set subjects contribute the standard score function of the true survival outcome. The non-validation set subjects contribute an estimated conditional expected score function of the true survival outcome given the observed

uncertain outcome. The applicability of such a method to survival outcomes is currently under investigation.

# APPENDIX A

## CHAPTER 2 SUPPLEMENTARY MATERIALS

### A.1. Development of Estimated Likelihood

Those in the validation set have both true and uncertain outcomes, so they contribute the joint probability distribution function, $P(X, \delta, X^*, \delta^*)$, to the likelihood. Those in the non-validation set only have the uncertain outcome available so they contribute the probability distribution function of the uncertain outcome only, $P(X^*, \delta^*)$. By re-writing the joint distribution using Bayes' formula, the full likelihood is given by

$$L = \prod_{i \in V} P(X_i, \delta_i) P(X_i^*, \delta_i^* | X_i, \delta_i) \prod_{j \in \bar{V}} P(X_j^*, \delta_j^*). \tag{A.1}$$

The distribution of the uncertain outcome can also be re-written by marginalizing the joint probability. The true outcome consists of both the true observed time and true event indicator, both of which are discrete, so we sum the joint distribution over all possible values of those variables to obtain the marginal distribution of the uncertain outcome. For an individual subject $j$,

$$P(X_j^*, \delta_j^*) = \sum_{k=1}^{K} \sum_{\delta=0}^{1} P(x_k, \delta, X_j^*, \delta_j^*). \tag{A.2}$$

We can re-write the joint distribution using Bayes' formula again to obtain

$$P(X_j^*, \delta_j^*) = \sum_{k=1}^{K} \sum_{\delta=0}^{1} P(x_k, \delta) P(X_j^*, \delta_j^* | x_k, \delta) \tag{A.3}$$

and our likelihood then becomes

$$L = \prod_{i \in V} P(X_i, \delta_i) P(X_i^*, \delta_i^* | X_i, \delta_i) \prod_{j \in \bar{V}} \sum_{k=1}^{K} \sum_{\delta=0}^{1} P(x_k, \delta) P(X_j^*, \delta_j^* | x_k, \delta). \tag{A.4}$$

To avoid having to specify or assume the form of the relationship between the true and uncertain endpoints, we estimate the conditional probabilities empirically and get an estimated likelihood,

$$\hat{L} = \prod_{i \in V} P(X_i, \delta_i) \hat{P}(X_i^*, \delta_i^* | X_i, \delta_i) \prod_{j \in \bar{V}} \sum_{k=1}^{K} \sum_{\delta=0}^{1} P(x_k, \delta) \hat{P}(X_j^*, \delta_j^* | x_k, \delta). \tag{A.5}$$

We estimate the conditional probability by first re-writing the expression using Bayes' formula. Since the validation set contains both true and uncertain endpoints, we use the validation set to empirically estimate each of the resulting probability distributions. For a subject $j$ in the non-validation set,

$$\begin{aligned} \hat{P}(X_j^*, \delta_j^* | x_k, \delta) &= \frac{\hat{P}(X_j^*, \delta_j^*, x_k, \delta)}{\hat{P}(x_k, \delta)} \\ &= \frac{\frac{1}{n_V} \sum_{i \in V} I(X_i^* = X_j^*, \delta_i^* = \delta_j^*, X_i = x_k, \delta_i = \delta)}{\frac{1}{n_V} \sum_{i \in V} I(X_i = x_k, \delta_i = \delta)}. \end{aligned} \tag{A.6}$$

Each of the empirical probabilities are proportions that consistently estimate the probability distributions. We can similarly use empirical probabilities to estimate the conditional probability in the validation set. By doing so, the conditional probability does not contain any parameters and can therefore be factored out of the likelihood. Our estimated likelihood to be maximized then takes the form

$$\hat{L} \propto \prod_{i \in V} P(X_i, \delta_i) \prod_{j \in \bar{V}} \sum_{k=1}^{K} \sum_{\delta=0}^{1} P(x_k, \delta) \frac{\frac{1}{n_V} \sum_{i \in V} I(X_i^* = X_j^*, \delta_i^* = \delta_j^*, X_i = x_k, \delta_i = \delta)}{\frac{1}{n_V} \sum_{i \in V} I(X_i = x_k, \delta_i = \delta)}. \tag{A.7}$$

Finally, for the marginal distribution of the true outcome, we use the same distribution as we would in a standard survival setting. For a subject in the validation set $i$,

$$P(X_i, \delta_i) = \{F(x_{k_i - 1}) - F(x_{k_i})\}^{\delta_i} F(x_{k_i})^{1-\delta_i} G(x_{k_i - 1})^{\delta_i} \{G(x_{k_i - 1}) - G(x_{k_i})\}^{1-\delta_i}. \tag{A.8}$$

In the validation set portion of the likelihood, the censoring distributions can also be factored out of the likelihood. However, the sum in the non-validation set portion of the likelihood prevents the

censoring distribution from being factored out. Our final estimated likelihood is

$$\begin{aligned}\hat{L} \propto & \prod_{i\in V}\{F(x_{k_i-1})-F(x_{k_i})\}^{\delta_i}F(x_{k_i})^{1-\delta_i} \\ & \cdot\prod_{j\in\bar{V}}\sum_{k=1}^{K}\sum_{\delta=0}^{1}\Bigg[\{F(x_{k-1})-F(x_k)\}^{\delta}F(x_k)^{1-\delta}G(x_{k-1})^{\delta}\{G(x_{k-1})-G(x_k)\}^{1-\delta} \\ & \qquad\cdot\frac{\frac{1}{n_V}\sum_{i\in V}I(X_i^*=X_j^*,\delta_i^*=\delta_j^*,X_i=x_k,\delta_i=\delta)}{\frac{1}{n_V}\sum_{i\in V}I(X_i=x_k,\delta_i=\delta)}\Bigg]. \end{aligned} \tag{A.9}$$

In the case of a perfect uncertain outcome, we assume $P(X,\delta|X^*,\delta^*)=1$. Then for all $j\in\bar{V}$, there exists a unique $(x_{k_j},\delta_j)$ such that $I(X_i^*=X_j^*,\delta_i^*=\delta_j^*,X_i=x_{k_j},\delta_i=\delta_j)>0$. Then the sum in the non-validation set becomes

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{\delta=0}^{1}\{F(x_{k-1})-F(x_k)\}^{\delta}F(x_k)^{1-\delta}G(x_{k-1})^{\delta}\{G(x_{k-1})-G(x_k)\}^{1-\delta}\hat{P}(X_j^*,\delta_j^*|x_k,\delta) \\ &\quad=\{F(x_{k_j-1})-F(x_{k_j})\}^{\delta_j}F(x_{k_j})^{1-\delta_j}G(x_{k_j-1})^{\delta_j}\{G(x_{k_j-1})-G(x_{k_j})\}^{1-\delta_j}\hat{P}(X_j^*,\delta_j^*|x_{k_j},\delta_j)\end{aligned} \tag{A.10}$$

and the estimated likelihood simplifies to

$$\begin{aligned}\hat{L} &\propto \prod_{i\in V}\{F(x_{k_i-1})-F(x_{k_i})\}^{\delta_i}F(x_{k_i})^{1-\delta_i} \\ &\quad\cdot\prod_{j\in\bar{V}}\{F(x_{k_j-1})-F(x_{k_j})\}^{\delta_j}F(x_{k_j})^{1-\delta_j}G(x_{k_j-1})^{\delta_j}\{G(x_{k_j-1})-G(x_{k_j})\}^{1-\delta_j}\hat{P}(X_j^*,\delta_j^*|x_{k_j},\delta_j) &\text{(A.11)}\\ &\propto \prod_{i\in V}\{F(x_{k_i-1})-F(x_{k_i})\}^{\delta_i}F(x_{k_i})^{1-\delta_i}\prod_{j\in\bar{V}}\{F(x_{k_j-1})-F(x_{k_j})\}^{\delta_j}F(x_{k_j})^{1-\delta_j} &\text{(A.12)}\\ &=\prod_{i=1}^{n}\{F(x_{k_i-1})-F(x_{k_i})\}^{\delta_i}F(x_{k_i})^{1-\delta_i}. &\text{(A.13)}\end{aligned}$$

This is the likelihood in a standard survival setting using only true outcomes in all subjects.

In the case of a useless uncertain outcome, we assume $P(X^*,\delta^*|X,\delta)=P(X^*,\delta^*)$. Then we can use an estimate of the marginal probability to estimate the conditional probability, or $\hat{P}(X^*,\delta^*|X,\delta)=$

$\hat{P}(X^*, \delta^*)$. Then the sum in the non-validation set becomes

$$\sum_{k=1}^{K}\sum_{\delta=0}^{1}\{F(x_{k-1}) - F(x_k)\}^{\delta}F(x_k)^{1-\delta}G(x_{k-1})^{\delta}\{G(x_{k-1}) - G(x_k)\}^{1-\delta}\hat{P}(X_j^*, \delta_j^*|x_k, \delta)$$
$$= \sum_{k=1}^{K}\sum_{\delta=0}^{1}\{F(x_{k-1}) - F(x_k)\}^{\delta_j}F(x_k)^{1-\delta_j}G(x_{k-1})^{\delta_j}\{G(x_{k-1}) - G(x_k)\}^{1-\delta_j}\hat{P}(X_j^*, \delta_j^*). \quad \text{(A.14)}$$

Because the estimated probability of the uncertain outcome does not contain any true outcome values, it can be factored out of the sum. The remaining sum then equals 1, since it is the sum of the probability distribution of the true outcome taken over all possible true outcome values. The estimated likelihood then simplifies to

$$\hat{L} \propto \prod_{i\in V}\{F(x_{k_i-1}) - F(x_{k_i})\}^{\delta_i}F(x_{k_i})^{1-\delta_i}\prod_{j\in\bar{V}}\hat{P}(X_j^*, \delta_j^*)$$
$$\propto \prod_{i\in V}\{F(x_{k_i-1}) - F(x_{k_i})\}^{\delta_i}F(x_{k_i})^{1-\delta_i}. \quad \text{(A.15)}$$

This is the likelihood in a standard survival setting using only true outcomes in the validation set.

## A.2. Correlation Calculation

Assume $T \sim$Unif$[1, 8]$, $T^* = T + \epsilon$, $\epsilon \sim$Unif$[0, \zeta]$, and $\epsilon$ is independent of $T$. Then the correlation between $T$ and $T^*$ can be written as

$$\rho = \frac{\text{Cov}(T, T^*)}{\sqrt{\text{Var}(T)\text{Var}(T^*)}}$$
$$= \frac{\text{Cov}(T, T+\epsilon)}{\sqrt{\text{Var}(T)\text{Var}(T+\epsilon)}}$$
$$= \frac{\text{Var}(T)}{\sqrt{\text{Var}(T)[\text{Var}(T) + \text{Var}(\epsilon)]}}$$
$$\Rightarrow \rho^2 = \frac{\text{Var}(T)^2}{\text{Var}(T)[\text{Var}(T) + \text{Var}(\epsilon)]}$$

We then solve for Var($\epsilon$) since this is the only part that contains $\zeta$.

$$\begin{aligned}
\Rightarrow \text{Var}(T)^2 &= \rho^2 \text{Var}(T)^2 + \rho^2 \text{Var}(T)\text{Var}(\epsilon) \\
\Rightarrow \text{Var}(\epsilon) &= \frac{\text{Var}(T)^2(1-\rho^2)}{\text{Var}(T)\rho^2} \\
&= \text{Var}(T)\frac{1+\rho^2}{\rho^2}
\end{aligned}$$

Finally, we substitute the variances of $T$ and $\epsilon$ using the variance formula for a discrete uniform random variable.

$$\begin{aligned}
\Rightarrow \frac{(\zeta+1)^2-1}{12} &= \frac{8^2-1}{12}\frac{1-\rho^2}{\rho^2} \\
\Rightarrow (\zeta+1)^2 &= (8^2-1)\frac{1-\rho^2}{\rho^2}+1 \\
\Rightarrow \zeta &= \sqrt{63\frac{1-\rho^2}{\rho^2}+1}-1
\end{aligned}$$

To ensure that $\zeta$ is an integer, we take the floor of the function as the maximum possible value for $\epsilon$.

## A.3. Description of the Alzheimer's Disease Neuroimaging Initiative

The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of ADNI is Michael W. Weiner, MD, VA Medical Center and University of California - San Francisco. ADNI is the result of efforts of many co- investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI

has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles.

## A.4. Supplementary Table 1

Table A.1: Simulation Results for Type 1 Censoring and $n = 500$

| $r$ | $\rho$ | Method | Bias $\times 10^{-3}$ | SD | $\widehat{\text{SE}}$ | MSE $\times 10^{-3}$ | RE | Cov |
|---|---|---|---|---|---|---|---|---|
| 25 | 0.01 | Proposed | 0.57 | 0.022 | 0.022 | 0.49 | 1.33 | 0.95 |
| | | Comp K-M | 0.34 | 0.022 | 0.022 | 0.49 | 1.33 | 0.95 |
| | | Naïve K-M | 498.02 | 0.002 | 0.002 | 309.79 | 0.01 | 0.00 |
| | 0.25 | Proposed | 0.59 | 0.022 | 0.022 | 0.49 | 1.32 | 0.96 |
| | | Comp K-M | 0.34 | 0.022 | 0.022 | 0.49 | 1.33 | 0.95 |
| | | Naïve K-M | 450.70 | 0.008 | 0.009 | 248.67 | 0.26 | 0.00 |
| | 0.50 | Proposed | 0.19 | 0.022 | 0.022 | 0.48 | 1.28 | 0.96 |
| | | Comp K-M | 0.34 | 0.022 | 0.022 | 0.49 | 1.33 | 0.95 |
| | | Naïve K-M | 385.52 | 0.013 | 0.013 | 175.92 | 0.59 | 0.00 |
| | 0.75 | Proposed | -0.15 | 0.021 | 0.021 | 0.45 | 1.21 | 0.96 |
| | | Comp K-M | 0.34 | 0.022 | 0.022 | 0.49 | 1.33 | 0.95 |
| | | Naïve K-M | 286.65 | 0.016 | 0.016 | 91.47 | 0.90 | 0.00 |
| | 1.00 | Proposed | -0.25 | 0.019 | 0.019 | 0.38 | 1.00 | 0.95 |
| | | Comp K-M | 0.34 | 0.022 | 0.022 | 0.49 | 1.33 | 0.95 |
| | | Naïve K-M | -0.24 | 0.019 | 0.019 | 0.38 | 1.00 | 0.95 |
| 50 | 0.01 | Proposed | 0.54 | 0.028 | 0.027 | 0.78 | 2.08 | 0.96 |
| | | Comp K-M | -0.26 | 0.028 | 0.027 | 0.78 | 2.08 | 0.95 |
| | | Naïve K-M | 498.02 | 0.002 | 0.002 | 309.79 | 0.01 | 0.00 |
| | 0.25 | Proposed | 1.00 | 0.028 | 0.027 | 0.79 | 2.11 | 0.96 |
| | | Comp K-M | -0.26 | 0.028 | 0.027 | 0.78 | 2.08 | 0.95 |
| | | Naïve K-M | 450.70 | 0.008 | 0.009 | 248.67 | 0.26 | 0.00 |
| | 0.50 | Proposed | -0.00 | 0.027 | 0.027 | 0.77 | 2.04 | 0.95 |
| | | Comp K-M | -0.26 | 0.028 | 0.027 | 0.78 | 2.08 | 0.95 |
| | | Naïve K-M | 385.52 | 0.013 | 0.013 | 175.92 | 0.59 | 0.00 |
| | 0.75 | Proposed | -0.39 | 0.025 | 0.025 | 0.65 | 1.75 | 0.96 |
| | | Comp K-M | -0.26 | 0.028 | 0.027 | 0.78 | 2.08 | 0.95 |
| | | Naïve K-M | 286.65 | 0.016 | 0.016 | 91.47 | 0.90 | 0.00 |
| | 1.00 | Proposed | -0.24 | 0.019 | 0.019 | 0.38 | 1.00 | 0.95 |
| | | Comp K-M | -0.26 | 0.028 | 0.027 | 0.78 | 2.08 | 0.95 |
| | | Naïve K-M | -0.24 | 0.019 | 0.019 | 0.38 | 1.00 | 0.95 |
| 75 | 0.01 | Proposed | 0.79 | 0.038 | 0.039 | 1.44 | 3.86 | 0.97 |
| | | Comp K-M | -1.93 | 0.037 | 0.038 | 1.39 | 3.74 | 0.96 |
| | | Naïve K-M | 498.02 | 0.002 | 0.002 | 309.79 | 0.01 | 0.00 |
| | 0.25 | Proposed | 8.77 | 0.041 | 0.041 | 1.75 | 4.66 | 0.97 |
| | | Comp K-M | -1.93 | 0.037 | 0.038 | 1.39 | 3.74 | 0.96 |
| | | Naïve K-M | 450.70 | 0.008 | 0.009 | 248.67 | 0.26 | 0.00 |
| | 0.50 | Proposed | 4.59 | 0.042 | 0.042 | 1.82 | 4.88 | 0.96 |
| | | Comp K-M | -1.93 | 0.037 | 0.038 | 1.39 | 3.74 | 0.96 |
| | | Naïve K-M | 385.52 | 0.013 | 0.013 | 175.92 | 0.59 | 0.00 |
| | 0.75 | Proposed | 1.87 | 0.038 | 0.038 | 1.47 | 4.04 | 0.96 |
| | | Comp K-M | -1.93 | 0.037 | 0.038 | 1.39 | 3.74 | 0.96 |
| | | Naïve K-M | 286.65 | 0.016 | 0.016 | 91.47 | 0.90 | 0.00 |
| | 1.00 | Proposed | -0.25 | 0.019 | 0.019 | 0.38 | 1.00 | 0.95 |
| | | Comp K-M | -1.93 | 0.037 | 0.038 | 1.39 | 3.74 | 0.96 |
| | | Naïve K-M | -0.24 | 0.019 | 0.019 | 0.38 | 1.00 | 0.95 |

$r$ is the percent missing and $\rho$ is the correlation between true and uncertain outcomes. Proposed refers to the proposed estimator, Comp K-M refers to the complete-case Kaplan-Meier estimator, and Naïve K-M refers to the naïve Kaplan-Meier estimator. SD is standard deviation of estimates across simulations, $\widehat{\text{SE}}$ is estimated standard error of the estimate, MSE is mean squared error, RE is relative efficiency, Cov is 95% coverage, all averaged across time.

## A.5. Supplementary Table 2

Table A.2: Simulation Results for Random Censoring and $n = 500$

| $r$ | $C$ | Method | Bias $\times 10^{-3}$ | SD | $\hat{SE}$ | MSE $\times 10^{-3}$ | RE | Cov |
|---|---|---|---|---|---|---|---|---|
| 25 | S | Proposed | 0.33 | 0.022 | 0.022 | 0.50 | 1.18 | 0.95 |
| | | Comp K-M | -0.72 | 0.024 | 0.024 | 0.58 | 1.36 | 0.95 |
| | | Naïve K-M | 119.47 | 0.019 | 0.019 | 14.86 | 0.83 | 0.00 |
| | L | Proposed | 0.24 | 0.025 | 0.024 | 0.63 | 1.18 | 0.95 |
| | | Comp K-M | -0.79 | 0.026 | 0.026 | 0.69 | 1.32 | 0.96 |
| | | Naïve K-M | 119.28 | 0.02 | 0.02 | 14.87 | 0.76 | 0.00 |
| 50 | S | Proposed | -0.19 | 0.025 | 0.027 | 0.64 | 1.51 | 0.96 |
| | | Comp K-M | -1.12 | 0.029 | 0.029 | 0.85 | 2.00 | 0.96 |
| | | Naïve K-M | 119.47 | 0.019 | 0.019 | 14.86 | 0.83 | 0.00 |
| | L | Proposed | -0.19 | 0.029 | 0.032 | 0.89 | 1.64 | 0.96 |
| | | Comp K-M | -0.29 | 0.033 | 0.032 | 1.12 | 2.12 | 0.95 |
| | | Naïve K-M | 119.28 | 0.020 | 0.020 | 14.87 | 0.76 | 0.00 |

$r$ is the percent missing and $C$ is the amount of censoring, where S means small (30%) and L means large (50%). Proposed refers to the proposed estimator, Comp K-M refers to the complete-case Kaplan-Meier estimator, and Naïve K-M refers to the naïve Kaplan-Meier estimator. SD is standard deviation of estimates across simulations, $\hat{SE}$ is estimated standard error of the estimate, MSE is mean squared error, RE is relative efficiency, Cov is 95% coverage, all averaged across time.

## A.6. Supplementary Table 3

Table A.3: Simulation Results for Data Missing at Random and $n = 500$

| Censoring | $\rho/C$ | Method | Bias $\times 10^{-3}$ | SD | $\widehat{\text{SE}}$ | MSE $\times 10^{-3}$ | RE | Cov |
|---|---|---|---|---|---|---|---|---|
| Type 1 | 0.01 | Proposed | 0.56 | 0.029 | 0.030 | 0.89 | 2.36 | 0.97 |
| | | Comp K-M | -1.32 | 0.029 | 0.030 | 0.88 | 2.34 | 0.96 |
| | | Naïve K-M | 498.02 | 0.002 | 0.002 | 309.79 | 0.01 | 0.00 |
| | 0.25 | Proposed | 13.93 | 0.030 | 0.031 | 0.92 | 2.46 | 0.97 |
| | | Comp K-M | -12.17 | 0.028 | 0.029 | 0.82 | 2.20 | 0.93 |
| | | Naïve K-M | 450.70 | 0.008 | 0.009 | 248.67 | 0.26 | 0.00 |
| | 0.50 | Proposed | 20.72 | 0.029 | 0.029 | 0.85 | 2.30 | 0.93 |
| | | Comp K-M | -25.75 | 0.027 | 0.028 | 0.76 | 2.03 | 0.85 |
| | | Naïve K-M | 385.52 | 0.013 | 0.013 | 175.92 | 0.59 | 0.00 |
| | 0.75 | Proposed | 15.75 | 0.026 | 0.026 | 0.70 | 1.92 | 0.92 |
| | | Comp K-M | -42.67 | 0.026 | 0.027 | 0.71 | 1.89 | 0.64 |
| | | Naïve K-M | 286.65 | 0.016 | 0.016 | 91.47 | 0.90 | 0.00 |
| | 1.00 | Proposed | -0.25 | 0.019 | 0.019 | 0.38 | 1.00 | 0.95 |
| | | Comp K-M | -21.32 | 0.025 | 0.025 | 0.65 | 1.70 | 0.80 |
| | | Naïve K-M | -0.24 | 0.019 | 0.019 | 0.38 | 1.00 | 0.95 |
| Random | S | Proposed | 2.43 | 0.025 | 0.026 | 0.62 | 1.44 | 0.95 |
| | | Comp K-M | -34.72 | 0.027 | 0.027 | 0.75 | 1.77 | 0.74 |
| | | Naïve K-M | 119.47 | 0.019 | 0.019 | 14.86 | 0.83 | 0.00 |
| | L | Proposed | 4.68 | 0.029 | 0.031 | 0.91 | 1.62 | 0.95 |
| | | Comp K-M | -43.16 | 0.031 | 0.030 | 0.95 | 1.90 | 0.69 |
| | | Naïve K-M | 119.28 | 0.020 | 0.020 | 14.87 | 0.76 | 0.00 |

Censoring is the type of the censoring mechanism and $\rho/C$ either represents the correlation $\rho$ between true and uncertain outcomes or represents the amount of censoring, where S means small (30%) and L means large (50%). Proposed refers to the proposed estimator, Comp K-M refers to the complete-case Kaplan-Meier estimator, and Naïve K-M refers to the naïve Kaplan-Meier estimator. SD is standard deviation of estimates across simulations, $\widehat{\text{SE}}$ is estimated standard error of the estimate, MSE is mean squared error, RE is relative efficiency, Cov is 95% coverage, all averaged across time.

# BIBLIOGRAPHY


K Andersen, L J Launder, M E Dewey, L Letenneur, A Ott, J Copeland, J-F Dartigues, P Kragh-Sorensen, M Baldereschi, C Brayne, A Lobo, J M Martinez-Lage, T Stijnen, A Hofman, and the EURODEM Incidence Research Group. Gender Differences in the Incidence of AD and Vascular Dementia: The EURODEM Studies. *Neurology*, 53(9):1992–1997, 1999.

R Balasubramanian and S W Lagakos. Estimation of the timing of perinatal transmission of HIV. *Biometrics*, 57(4):1048–58, December 2001. ISSN 0006-341X.

Steven T DeKosky, Annette Fitzpatrick, Diane G Ives, Judith Saxton, Jeff Williamson, Oscar L Lopez, Gregory Burke, Linda Fried, Lewis H Kuller, John Robbins, Russell Tracy, Nancy Woolard, Leslie Dunn, Richard Kronmal, Richard Nahin, and Curt Furberg. The Ginkgo Evaluation of Memory (GEM) study: design and baseline data of a randomized trial of Ginkgo biloba extract in prevention of dementia. *Contemporary clinical trials*, 27(3):238–53, June 2006. ISSN 1551-7144. doi: 10.1016/j.cct.2006.02.007. URL http://www.ncbi.nlm.nih.gov/pubmed/16627007.

Bradley Efron. The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association1*, 72(359):557–565, 1977.

Thomas R Fleming, Ross L Prentice, Margaret S Pepe, and David Glidden. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in Medicine*, 13:955–968, 1994.

L. Fratiglioni, M. Viitanen, E. von Strauss, V. Tontodonati, A. Herlitz, and B. Winblad. Very Old Women at Highest Risk of Dementia and Alzheimer's Disease: Incidence Data from the Kungsholmen Project, Stockholm. *Neurology*, 48(1):132–138, January 1997. ISSN 0028-3878. doi: 10.1212/WNL.48.1.132.

L.S. Freedman. Tables of the number of patients required in clinical trials using the logrank test. *Statistics in medicine*, 1:121–129, 1982. ISSN 0277-6715.

Frank E Harrell Jr. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis, 2001.

Clifford R Jack Jr, David S Knopman, William J Jagust, Leslie M Shaw, Paul S Aisen, Michael W Weiner, Ronald C Petersen, and John Q Trojanowski. Hypothetical Model of Dynamic Biomarkers of the Alzheimer's Pathological Cascade. *Lancet Neurology*, 9(1):119–128, 2010. doi: 10.1016/S1474-4422(09)70299-6.Hypothetical.

John P. Klein and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, New York, NY, 2nd ed. edition, 2003. ISBN 038795399X.

E Lakatos. Sample size determination in clinical trials with time-dependent rates of losses and noncompliance. *Controlled clinical trials*, 7(3):189–99, September 1986. ISSN 0197-2456. URL http://www.ncbi.nlm.nih.gov/pubmed/3802834.

E Lakatos. Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics*, 44(1): 229–41, March 1988. ISSN 0006-341X. URL http://www.ncbi.nlm.nih.gov/pubmed/3358991.

L Letenneur, V Gilleron, D Commenges, C Helmer, J M Orgogozo, and J F Dartigues. Are sex and educational level independent predictors of dementia and Alzheimer s disease ? Incidence data from the PAQUID project. *Journal of Neurology, Neurosurgery, & Psychiatry*, 66:177–183, 1999.

Joan Lindsay, Danielle Laurin, René Verreault, Réjean Hébert, Barbara Helliwell, Gerrry B. Hill, and Ian McDowell. Risk Factors for Alzheimer's Disease: A Prospective Analysis from the Canadian Study of Health and Aging. *American Journal of Epidemiology*, 156(5):445–453, September 2002. ISSN 00029262. doi: 10.1093/aje/kwf074.

Amalia S Magaret. Incorporating validation subsets into discrete proportional hazards models for mismeasured outcomes. *Statistics in Medicine*, 27:5456–5470, 2008. doi: 10.1002/sim.

Amalia S Meier, Barbra A Richardson, and James P Hughes. Discrete Proportional Hazards Models for Mismeasured Outcomes. *Biometrics*, 59(4):947–954, 2003.

Michelle M Mielke, Vemuri Prashanthi, and Walter A Rocca. Clinical epidemiology of Alzheimer s disease: assessing sex and gender differences. *Clinical Epidemiology*, 6:37–48, 2014.

Peter T Nelson, Irina Alafuzoff, Eileen H Bigio, Constantin Bouras, Heiko Braak, Nigel J Cairns, Rudolph J Castellani, Barbara J Crain, Peter Davies, Kelly Del Tredici, Charles Duyckaerts, Matthew P Frosch, Vahram Haroutunian, Patrick R Hof, Christine M Hulette, Bradley T Hyman, Takeshi Iwatsubo, Kurt a Jellinger, Gregory a Jicha, Enikö Kövari, Walter a Kukull, James B Leverenz, Seth Love, Ian R Mackenzie, David M Mann, Eliezer Masliah, Ann C McKee, Thomas J Montine, John C Morris, Julie a Schneider, Joshua a Sonnen, Dietmar R Thal, John Q Trojanowski, Juan C Troncoso, Thomas Wisniewski, Randall L Woltjer, and Thomas G Beach. Correlation of Alzheimer Disease Neuropathologic Changes with Cognitive Status: A Review of the Literature. *Journal of neuropathology and experimental neurology*, 71(5):362–81, May 2012. ISSN 1554-6578. doi: 10.1097/NEN.0b013e31825018f7.

Alewijn Ott, Monique MB Breteler, Frans van Harskamp, Theo Stijnen, and Albert Hofman. Incidence and Risk of Dementia: The Rotterdam Study. *American Journal of Epidemiology*, 147(6): 574–580, 1998.

Margaret Sullivan Pepe, Marie Reilly, and Thomas R. Fleming. Auxiliary outcome data and the mean score method. *Journal of Statistical Planning and Inference*, 42(1-2):137–160, November 1994. ISSN 03783758. doi: 10.1016/0378-3758(94)90194-5. URL `http://linkinghub.elsevier.com/retrieve/pii/0378375894901945`.

MS Pepe. Inference using surrogate outcome data and a validation sample. *Biometrika*, 79(2): 355–365, 1992.

C Qiu, L Bäckman, B Winblad, H Agüero-Torres, and L Fratiglioni. The influence of education on clinically diagnosed dementia incidence and mortality data from the Kungsholmen Project. *Archives of neurology*, 58(12):2034–9, December 2001. ISSN 0003-9942.

B a Richardson and J P Hughes. Product limit estimation for infectious disease data when the diagnostic test for the outcome is measured with uncertainty. *Biostatistics*, 1(3):341–54, September 2000. ISSN 1465-4644. doi: 10.1093/biostatistics/1.3.341.

Christine Sattler, Pablo Toro, Peter Schönknecht, and Johannes Schröder. Cognitive activity, education and socioeconomic status as preventive factors for mild cognitive impairment and Alzheimer's disease. *Psychiatry research*, 196(1):90–5, March 2012. ISSN 0165-1781. doi: 10.1016/j.psychres.2011.11.012.

David Schoenfeld. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, 68(1):316–319, 1981.

David A Schoenfeld. Sample-Size Formula for the PH Regression Model. *Biometrics*, 39(2):499–503, 1983.

LM Shaw, Hugo Vanderstichele, Malgorzata Knapik-Czajka, Christopher M Clark, Paul S Aisen, Ronald C Petersen, Kaj Blennow, Holly Soares, Adam Simon, Piotr Lewczuk, Robert Dean, Eric Siemers, William Potter, Virginia M-Y Lee, John Q Trojanowski, and Alzheimer's DIsease Neuroimaging Initiative. Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Annals of Neurology*, 65(4):403–413, 2009. doi: 10.1002/ana.21610. Cerebrospinal.

J H Shih. Sample size calculation for complex clinical trials with survival endpoints. *Controlled clinical trials*, 16(6):395–407, December 1995. ISSN 0197-2456. URL `http://www.ncbi.nlm.nih.gov/pubmed/8720017`.

B. W. Silverman. *Density Estimation.* Chapman and Hall, London, 1992.

Jeffrey S. Simonoff. *Smoothing Methods in Statistics*. Springer, New York, NY, 1996.

S M Snapinn. Survival analysis with uncertain endpoints. *Biometrics*, 54(1):209–18, March 1998. ISSN 0006-341X.

Yaakov Stern, Barry Gurland, Thomas K Tatemichi, David Wilder, and Richard Mayeux. Influence of Education and Occupation on the Incidence of Alzheimer ' s Disease. *Journal of the American Medical Association*, 271(13):1004–1010, 1994.

Lesley a Stevens, Josef Coresh, Tom Greene, and Andrew S Levey. Assessing kidney function–measured and estimated glomerular filtration rate. *The New England journal of medicine*, 354 (23):2473–83, June 2006. ISSN 1533-4406. doi: 10.1056/NEJMra054415.

Michael W Weiner, Dallas P Veitch, Paul S Aisen, Laurel a Beckett, Nigel J Cairns, Robert C Green, Danielle Harvey, Clifford R Jack, William Jagust, Enchi Liu, John C Morris, Ronald C Petersen, Andrew J Saykin, Mark E Schmidt, Leslie Shaw, Judith a Siuciak, Holly Soares, Arthur W Toga, and John Q Trojanowski. The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. *Alzheimer's & Dementia*, 8(1 Suppl):S1–68, February 2012. ISSN 1552-5279. doi: 10.1016/j.jalz.2011.09.172.

Sevil Yasar, Jin Xia, Wenliang Yao, Curt D Furberg, Qian-li Xue, Carla I Mercado, Annette L Fitzpatrick, Linda P Fried, Claudia H Kawas, Kaycee M Sink, Jeff D Williamson, Steven T Dekosky, and Michelle C Carlson. Antihypertensive drugs decrease risk of Alzheimer disease. *Neurology*, 81:896–903, 2013.